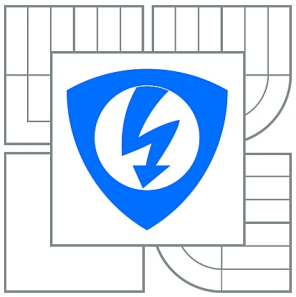
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF TELECOMMUNICATIONS

IDENTIFIKACE ŘEČOVÉ AKTIVITY V RUŠENÉM ŘEČOVÉM SIGNÁLU

IDENTIFICATION OF SPEECH ACTIVITY IN NOISY SPEECH SIGNAL

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

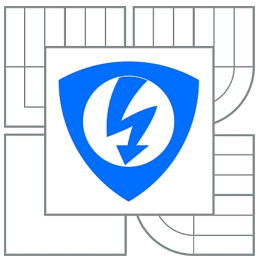
AUTOR PRÁCE — Bc. MARTIN PELIKÁN

AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. ZDENĚK SMÉKAL, CSc.

SUPERVISOR

BRNO 2013



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student:	Bc. Martin Pelikán	***ID:***	120610
Ročník:	2	***Akademický rok:***	2012/2013

NÁZEV TÉMATU:

Identifikace řečové aktivity v rušeném řečovém signálu

POKYNY PRO VYPRACOVÁNÍ:

Základním problémem metod zvýraznění řeči je úplné oddělení přirozeného šumu, který vzniká při správně artikulaci znělých (sonorů) a neznělých souhlásek (konsonant) od šumu a rušení okolního prostředí. Cílem diplomové práce je najít efektivní metodu, která by dokázala věrohodně identifikovat a oddělit nežádoucí šum a rušení od toho, který do řeči patří. Řešení problému souvisí s identifikací pauz bez řečové aktivity, v nichž je možné identifikovat vlastnosti šumu a rušení. Jakmile je správně šum určen, pak již je možné využít různých metod pro jeho odstranění. Navržená efektivní metoda by měla být testována v programu Matlab.

DOPORUČENÁ LITERATURA:

[1] SMÉKAL, Z.: Číslicové zpracování řeči (MZPR). Elektronické učební texty pro magisterské studium, VUT Brno, 2011.
[2] PSUTKA, J., MULLER, L., MATOUŠEK, J., RADOVÁ, V.: Mluvíme s počítačem česky. Academia, Praha 2006. ISBN 80-2100-1309-1
[3] KRČMOVÁ, M.: Fonetika. Elektronické texty. MU Brno 2003. http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html

Termín zadání:	11.2.2013	***Termín odevzdání:***	29.5.2013

Vedoucí práce: prof. Ing. Zdeněk Smékal, CSc.
Konzultanti diplomové práce:

prof. Ing. Kamil Vrba, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

SIX research centre
sensor, information and communication systems

Faculty of Electrical Engineering
and Communication

Brno University of Technology
Purkynova 118, CZ-61200 Brno, Czechia

http://www.six.feec.vutbr.cz

Výzkum popsaný v této habilitační práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

2007–13
OP Výzkum a vývoj pro inovace

ABSTRAKT

Diplomová práce se zabývá hledáním optimálního nastavení parametrů detektorů řečové aktivity a následnou filtrací neužitečného šumu ze signálu. Nejprve jsou teoreticky popsány metody zpracování signálu, poté jednotlivé detektory řečové aktivy spolu s metodami filtrování šumu a nakonec porovnání výsledků filtrace šumu pro různá nastavení VAD a přesnost detekce řeč /pauza.

KLÍČOVÁ SLOVA

VAD, kepstrální detektor, zvýraznění, šum, pauza, identifikace, spektrální odečítání

ABSTRACT

This thesis is focused on finding the optimal set of parameters of voice activity detectors and following filtering of the noise from the signal. Firstly the signal processing methods are theoretically defined, then voice activity detectors with noise filtering methods are described and in the end the results of noise filtering and the accuracy of the speech / pause detection for various settings are presented.

KEYWORDS

VAD, cepstral detector, enhancement, noise, pause, identification, spectral subtraction

PELIKÁN, M. *Identifikace řečové aktivity v rušeném řečovém signálu* . Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2013. 62 s. Vedoucí semestrální práce prof. Ing. Zdeněk Smékal, CSc..

PROHLÁŠENÍ

Prohlašuji, že svoji diplomovou práci na téma „Identifikace řečové aktivity v rušeném řečovém signálu“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení §11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

(podpis autora)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu prof. Ing. Zdeňku Smékalovi, CSc. za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé diplomové práce.

Brno ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Dnešní doba je doba technologického pokroku. Náš život je stále více propojován s výpočetní technikou. Mobilní telefony, internet, počítače. Všechno je na dosah ruky. S tím také roste míra používání těchto zařízení. Vezměme si například takové mobilní telefony. Podle údajů ČTÚ z roku připadá na každých sto obyvatel 135 SIM karet a mobilních telefonů. To je hodně. Navíc máme mobilní telefon stále k dispozici a voláme odkudkoli komukoli. Sice je to praktické a vlastně i výhodné, ale dochází k jedné nepříjemné věci. To je znehodnocování hovoru šumem. Je jedno, jestli se jedná o aditivní šum, vznikající v mikrofonu při nahrávání hlasu, či o šum konvoluční, získaný například hlukem z ulice. Každý šum kazí výsledný akustický projev, rozptyluje posluchače a dokáže zvuk znehodnotit až do takové míry, že bude neposlouchatelný. Dalším problémem je zvýšená režie při přenosu, kdy musíme přenášet i šum. Toto se netýká samozřejmě pouze mobilní komunikace. Můžeme sem zařadit i rozpoznávání hlasových povelů počítačem, televizní přenosy a podobně. Proto se snažíme šum nějakým způsobem filtrovat. K tomu existuje několik metod. Jednou z nich je detekovat v řečovém signálu úseky pauzy a řeči pomocí zařízení nazvaném „detektor řečové aktivity“. Když jsme schopni zjistit úseky neobsahující řeč, můžeme říct, že obsahují šum. Posléze jsme schopni vhodnými metodami šum odfiltrovat.

Tato diplomová práce se věnuje detektorům řečové aktivity, detekci řeč/ticho, následné optimalizaci detekčního procesu pomocí hledání optimálního nastavení detektoru a metodám odečítání šumu ze signálu kontaminovaného šumem. První část dokumentu tvoří teoretické informace o vzniku řeči a popisu šumu. Věnuje se také předzpracování řečového signálu, nutnému k dalšímu zpracování detektory řečové aktivity. Druhá kapitola je zaměřena na samotné detektory; jejich vlastnosti, využití, a funkci. Třetí kapitola popisuje možné metody odstranění šumu ze záznamu a podrobněji se zabývá metodou spektrálního odečítání. Poslední část čistě popisuje testovací metodiku, použití testovacích databází a prezentaci výsledků.

Závěrem jsou prezentovány výsledky praktické části, včetně osobních poznámek autora k celému experimentu.

1 ZPRACOVÁNÍ ŘEČOVÉHO SIGNÁLU

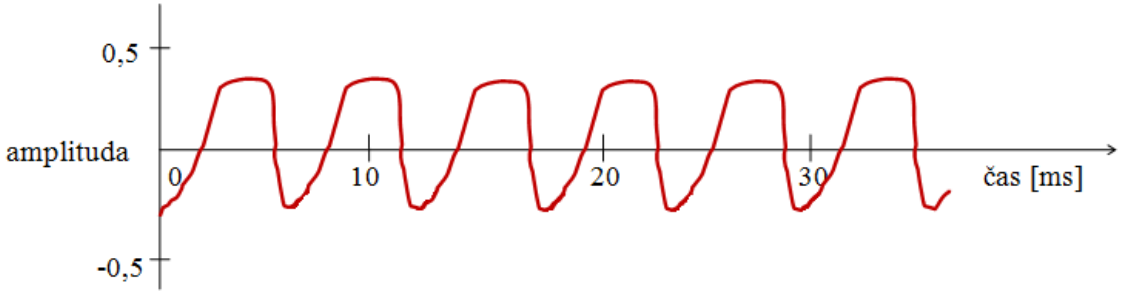
Zpracování signálu řeči je základním stavebním kamenem všech programů pro rozpoznávání a zvýrazňování řeči, detektorů řečové aktivity, kodérů pro přenos řečového signálu a dalších. Abychom byli s to pochopit, jak zpracování řečového signálu probíhá, musíme se ponořit do teorie tvorby řeči. Poté definujeme pojem šum a nakonec se budeme zabývat zpracováním signálu.

1.1 Tvorba řeči

Mluvená řeč je soubor akustických vln, tzv. akustického signálu, procházejícího komunikačním kanálem. Komunikačním kanálem se v tomto případě rozumí prostředí, kterým řeč putuje ze svého zdroje, z dýchacího ústrojí řečníka, až k příjemci, tedy uším posluchače. Akustický signál je mechanické vlnění o frekvencích slyšitelných člověkem. Rozsah lidského sluchu je v rozmezí 16 Hz až 20 kHz [10].

Akustický signál vzniká v dechovém ústrojí člověka. Jako zdroj energie slouží plíce při výdechu. Mozek vyšle impulz svalům dechového ústrojí a ty se podle požadavků mozku potřebným způsobem pohybují. Jejich pohybem se mění akustický tlak okolního vzduchu a vzniká tak akustická vlna.

Akustická vlna postupuje z plic přes průdušky a průdušnice do hlasového ústrojí, uloženého v hrtanu. Hlasové ústrojí slouží k vytvoření základního hlasového tónu, jehož úpravami vzniká mluvní hlas [12]. Nejdůležitější částí ústrojí jsou hlasivky, dva hlasové vazy pokryté sliznicí vedoucí napříč hrtanem. Jejich délka se pohybuje mezi 13mm až 15mm v závislosti na pohlaví řečníka. Hlasivky při průchodu akustickou vlnou začnou kmitat, přitom se prudce otvírají a zavírají. V tomto momentě vzniká kvaziperiodický budící signál mající tónový charakter. Označuje se termínem základní tón a představuje nosný zvuk řeči. Obr. 1.1 zobrazuje průběh základního tónu změřeného pomocí EGG. Z grafu je patrné, že vlastnosti hlasového traktu zůstávají neměnné v časovém rozmezí 10 až 20 ms.



Obr. 1.1: Ukázka hlasivkového signálu změřeného pomocí EGG

Hlasivky pracují dvěma způsoby a podle toho se vytváří rozdílné hlásky:

- rychlé kmitání hlasivek, vznik souhlásek a znělých samohlásek: a, á, e, é , i, í, o, ó, u, ú
- hlasivky jsou klidně rozevřeny a vznikají hlásky: t, ť, s, š, č, p, k, f

K vytvarování akustického signálu, procházejícího přes hlasové ústrojí, do výsledné podoby slouží artikulační ústrojí. Skládá se z nadhrtanových dutin (pasivní prvek) a artikulačních orgánů (aktivní prvek).

Důležitou částí samotné řeči jsou také pauzy, tvořící hraniční body promluvy. Pauzy ovlivňují jak plynulost řeči, tak i její přirozenost. Délky pauz se liší a to jestli oddělují větné celky, nebo pouze větné úseky. V tomto případě hovoříme o tzv. tiché pauze. Někdy může být pauza vyplněna např. vydechnutím, nadechnutím, či zakašláním řečníka a systémy detekce řeči / pauzy mohou daný úsek nesprávně označit jako řeč. Přesné vymezení pauzy a řeči může být v mnoha případech velice problematické.

1.2 Šum

Důvodů pro vznik detektorů hlasové aktivity bylo víc, jedním z nich je odfiltrování šumu a hluku v řečovém kanále, přenášejícím se např. přes telekomunikační kanály. Šum a hluk zhoršují zvukovou kvalitu hlasového hovoru, stejně jako zvyšuje datový tok komunikace ve VOIP telefonii a podobně. Hluk chápeme jako úzkopásmový až středně širokopásmový nestacionární signál, šum jako širokopásmový parazitní signál. Společně je budeme nazývat šum. Dělí se na aditivní a konvoluční.

Aditivní šum je skupina šumů, které se k užitečnému signálu přičtou v akustické rovině při záznamu akustickými snímači. Tento šum je nekorelovaný s řečovým signálem. Můžeme ho rozdělit na bílý a barevný nebo na stacionární a nestacionární.

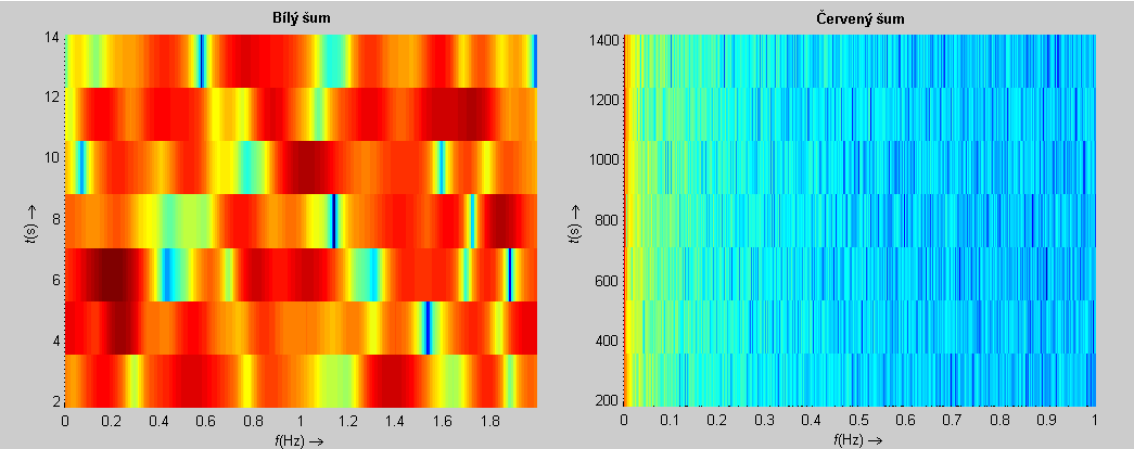
Konvoluční šum souvisí s užitečným signálem. Nejčastěji vzniká už při zachycení zvuku odrazem od okolních překážek, dále v průběhu přenosové cesty jako zkreslení způsobené odrazy nepřizpůsobeným vedením nebo časovým zpožděním, anebo na nelineárních aktivních prvcích. Tento druh zkreslení, pokud už vznikne, se odstraňuje jinými metodami [7].

Aditivní bílý Gaussovský šum (White Gaussian Noise) je náhodný nekorelovaný stacionární signál, nenesoucí žádnou informaci, s rovnoměrnou výkonovou spektrální hustotou. Je popsán rozptylem $\sigma^2[n]$ a střední hodnotou $\mu[n]$. Spektrální výkonová hustota bílého šumu je konstantní pro $\mu[n] = 0$ a platí, že $P[n] = \sigma^2[n]$. Bílý šum se v reálném prostředí nevyskytuje, avšak se využívá v modelech reálných systémů nebo např. v hudební technice.

Barevný aditivní šum má jiné rozložení energií ve spektru. Je aproximací bílého šumu a získá se jeho filtrací. Na Obr. 1.2 jsou spektra bílého a červeného[1] šumu.

Stacionární šum má výkonovou spektrální hustotu téměř konstantní v čase a bývá způsoben například zvukem větráku počítače či klimatizace, hlukem neakcelerujícího automobilového motoru, zvukem ještě šumem vzdálené konverzace apod.

Nestacionární šum, jehož spektrální charakteristiky se zřetelně mění v čase, vzniká například houkáním projíždějící sanitky, zvukem stisknutí klávesy na klávesnici, zvukem zvonku, či hlasitým dýcháním při mluvení, smíchem apod. [10].

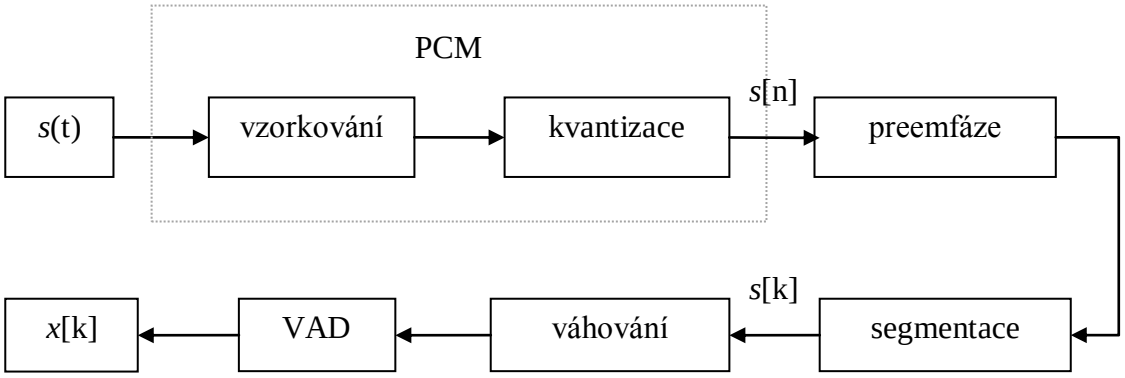


Obr. 1.2: Ukázka spektra bílého šumu (vlevo) a červeného šumu (vpravo)

[1] Červený šum se někdy označuje také jako hnědý nebo růžový

1.3 Předzpracování signálu

Abychom mohli využít detektoru řečové aktivity k zvýraznění řeči, musí být testovaný signál nejprve převeden z analogové podoby do digitální. Signál projde procesem zvýraznění vyšších frekvencí, rozdělí se na menší segmenty, jednotlivé segmenty jsou váhovány okenní funkcí a takto předzpracovaný signál se analyzuje metodami krátkodobé analýzy v bloku VAD, viz. Obr. 1.3. Výsledkem je označený signál $x[k]$.



Obr. 1.3: Blokové schéma zpracování vstupního signálu s(t)

1.3.1 Pulzní kódová modulace (PCM)

Akustický signál mluvené řeči je obvykle nahráván mikrofonem. Pro další zpracování signálu je nutno tyto analogové kmity převést do digitální podoby. Spojitý analogový signál nabývá nekonečně hodnot. My ale potřebujeme mít diskrétní digitální signál s konečným počtem hodnot. Převod z analogového do digitálního signálu se zajistí pulzní kódovou modulací. PCM se sestává ze dvou částí - vzorkování a kvantování.

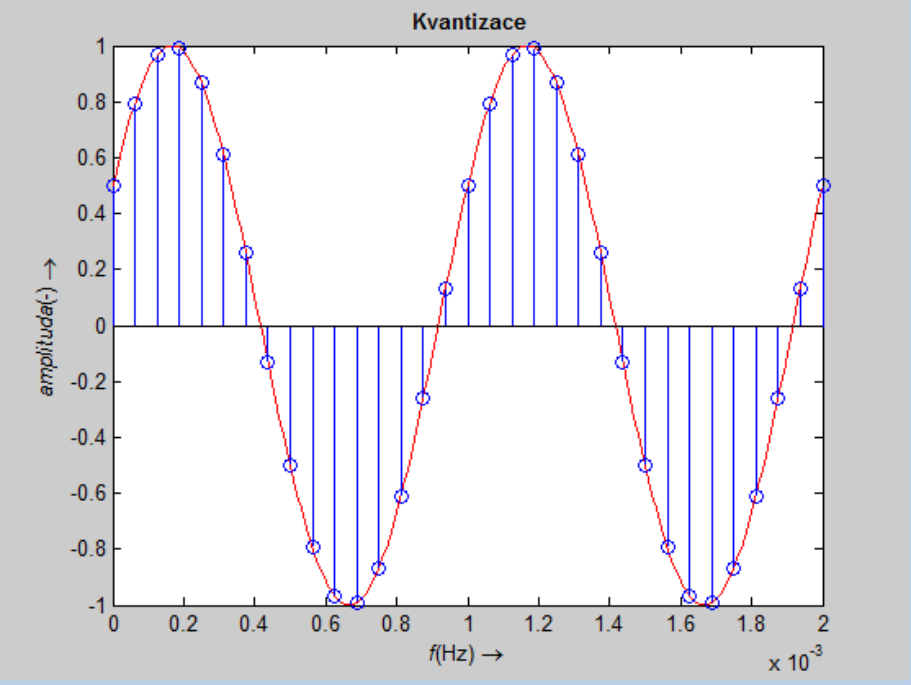
1.3.2 Vzorkování

Vzorkování se provádí tak, že se časová osa signálu rozdělí na stejné úseky a z každého úseku se vezme jeden vzorek. Je to proces transformace spojitého signálu $s(t)$ na posloupnost vzorků $s[n] = s(nT)$ diskrétních v čase. Vzorkování signálu probíhá v časových úsecích $t[n] = nT$, kde T je perioda vzorkování a n je z oboru přirozených čísel [10]. Vzorkovací frekvence f_v musí být zvolena alespoň dvakrát větší, než je frekvenční šířka pásma vzorkovaného signálu. Vyplívá to ze vzorkovacího teorému[2] $f_v \geq f_m$. Při nedodržení vzorkovacího teorému dojde k aliasingu, zkreslení složek vyšších frekvencí.

[2] Vzorkovací teorém, označovaný také jako Nyquistův teorém, pojmenovaný podle svého objevitele, Harryho Nyquista.

1.3.3 **Kvantizace a kódování**

Kvantizace je proces, při kterém hodnotě vzorku z nekonečné množiny přidělíme hodnotu z množiny konečné. Vzorkovač vyhodnocuje vzorky a hodnotu převádí podle předem definované kvantizační úrovně. S rostoucím počtem kvantizačních úrovní narůstá přesnost kvantování. Počet úrovní se volí ve tvaru 2^n.



Obr. 1.4: Proces kvantizace

1.3.4 **Preemfáze**

Preemfáze je proces zdůrazňování amplitud spektrálních složek signálu řeči s jejich narůstající frekvencí. Aplikuje se před vlastním zpracováním signálu číslicovým filtrem typu horní propust (FIR), zařazený za blok vzorkování a kvantizace. Je popsaný přenosovou funkcí

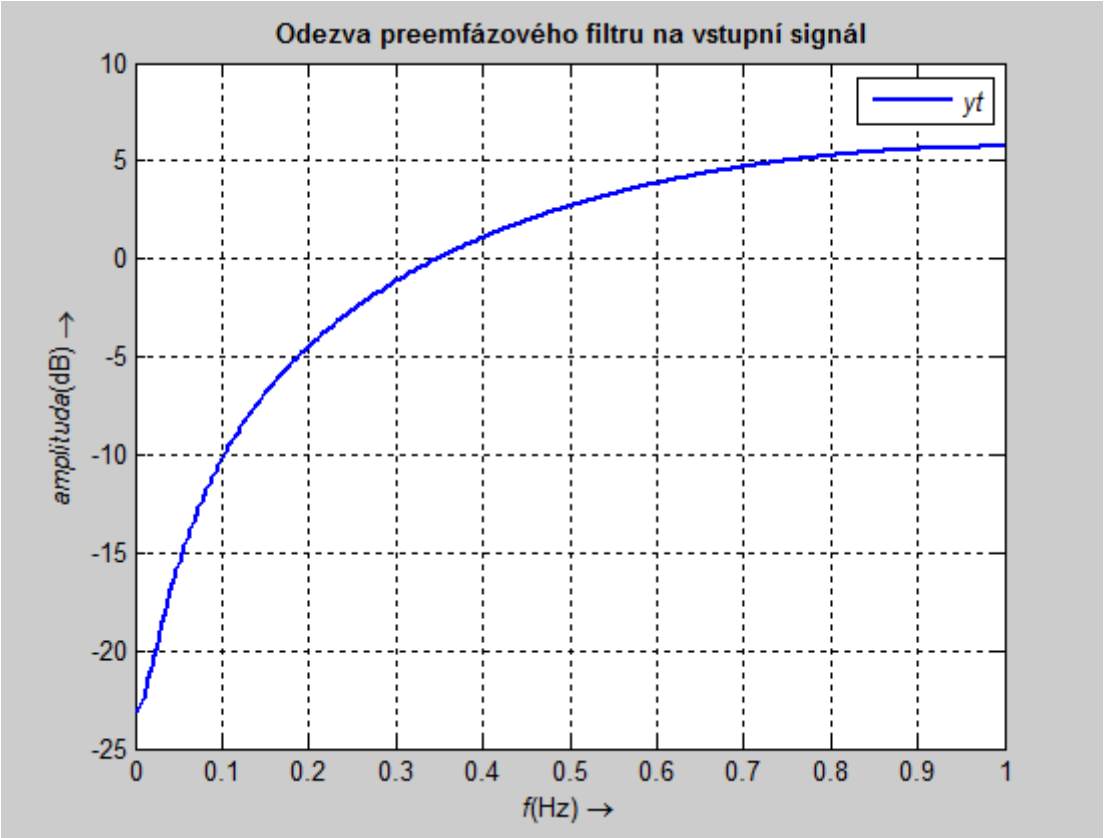
$$H(z) = 1 - a_1 \cdot z^{-1}. \qquad (1.1)$$

Koeficient a_1 nabývá hodnot 0,9 až 1. Diferenční rovnice odpovídající přenosové funkci $H(z)$ má tvar

$$y[n] = x[n] - a_1 \cdot x[n-1], \qquad (1.2)$$

kde $x[n]$ je vstupní vzorek filtru v čase n, $y[n]$ je výstup. Využití preemfáze je dáno vlastnostmi hlasového ústrojí, kdy citlivost lidského sluchu klesá se vzrůstající frekvencí signálu a zároveň klesají amplitudy spektrálních složek řečového signálu na vyšších frekvencích. Zdůrazňováním amplitud tento pokles do značné míry

kompenzujeme. Díky tomu dojde k relativnímu vyrovnání spektra přenášeného pásma, tedy amplitudy všech složek budou dosahovat přibližně (řádově) stejných úrovní. Při akustické rekonstrukci takto zpracovaného signálu musíme samozřejmě aplikovat inverzní filtraci [10]. Na Obr. 1.5 je zobrazeno zvýraznění amplitudy při použití preemfáze.



Obr. 1.5: Modulová charakteristika číslicového filtru

1.3.5 **Segmentace**

Podmínka při zpracování řečového signálu je taková, aby zpracovaný signál byl stacionární. Toho nelze dosáhnout na celém úseku signálu. Proto se signál musí rozdělit na *N* stejně dlouhých úseků, tzv. segmentů. Předpokládáme, že se vlastnosti hlasového traktu při promluvě v časovém úseku 10 až 30 ms nemění dostatečně rychle a zůstávají přibližně konstantní. Hovoříme o signálu jako o kvazistacionárním. Velice často se v praxi využívá velikost segmentů 16 ms.

1.3.6 **Váhování**

Dělení delšího úseku řečového signálu na krátké úseky je prováděno pomocí časových oken, aby nedocházelo k nepřesnosti ve zpracování. Ta by totiž mohla nastat při nenulovém překrytí rámců. Při zpracování řeči se nejvíce používá pravoúhlé

a Hammingovo okno.

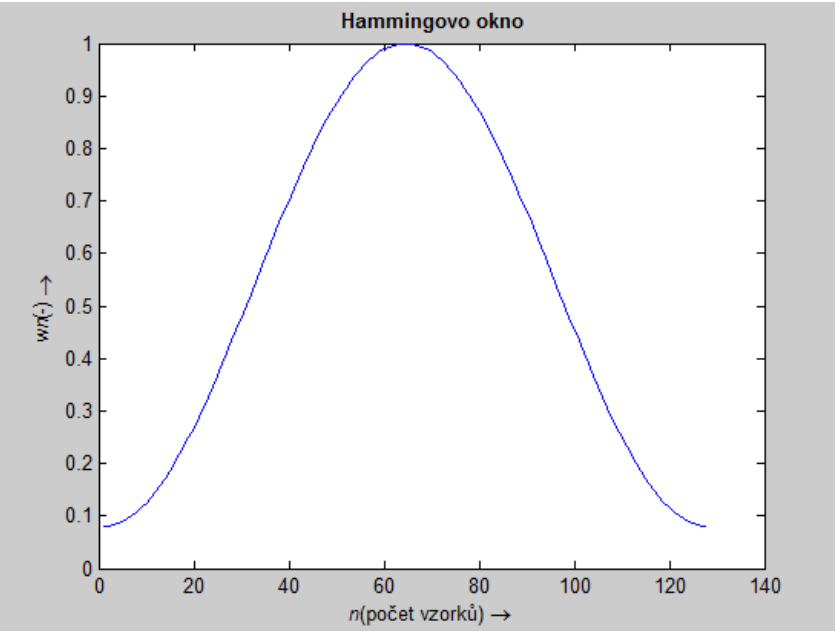
Pravoúhlé okno je definováno:

$$w[n] = 1, \text{ pro } n = 0, 1, \ldots, N\text{-}1,$$
$$w[n] = 0, \text{ pro všechna ostatní } n.$$

Hammingovo okno je definováno:

$$w[n] = 0{,}56 - 0{,}46 \cdot \cos\left[n \frac{2 \cdot \pi}{N} \right], \text{ pro } n = 0, 1, \ldots, N\text{-}1,$$
$$w[n] = 0, \text{ pro všechna ostatní } n.$$

Přestože pravoúhlé okno je jednodušší, často se používá Hammingovo okno, protože potlačuje vzorky na okrajích segmentů, čímž se zvyšuje stabilita některých výpočtů. Dále se Hammingovo okno používá především pro schopnost potlačit postranní laloky ve spektru, které nám u obdélníkového okna výrazně zkreslují skutečnou podobu spektra [7]. Vzhledem k tomu je potřeba zvolit míru překrytí vážených segmentů, nejčastěji se volí hodnota 50 %.



Obr. 1.6: Časový průběh Hammingova okna

1.4 Příznaky řeči

Řečové příznaky charakterizují základní vlastnosti řečového signálu. Jsou to informace extrahovány nejčastěji z krátkých úseků řečového signálu a mohou být reprezentovány jako číslo, vektor nebo dokonce dvojrozměrná matice. Délka těchto mikrosegmentů většinou bývá 10 ms. Výsledkem je časová posloupnost skupiny čísel.

Mezi často používané řečové příznaky patří:

- Základní tón F_0
- Formanty
- Lineární predikční koeficienty
- Melovské kepstrální koeficienty
- Dynamické koeficienty

1.4.1 Základní tón

Frekvence kmitů hlasivek F_0 se pohybuje mezi 60 Hz a 400 Hz. Je označována jako základní tón řeči a liší se v závislosti na věku a pohlaví osoby. Průběh základního tónu udává tzv. melodii řeči. Při běžné řeči se hodnota F_0 pohybuje zhruba v rozmezí jedné oktávy. Při zpěvu se rozsah zvětšuje a může přesahovat až 2 oktávy. Vztah základního tónu k základní periodě je popsán následovně:

$$F_0 = \frac{1}{T_0} = \frac{f_{vz}}{L}\,[\mathrm{Hz}], \tag{1.3}$$

kde T_0 je základní perioda, L je základní perioda ve vzorcích a f_{vz} je vzorkovací frekvence signálu.

Podle [1] nalezne základní tón uplatnění při:

- kódování řeči pomocí lineární predikce
- syntéze řeči a modelování prozodie
- analýze řeči
- detekci řečové aktivity v signálu
- identifikaci mluvčího

1.4.2 Formanty

Rezonanční vrcholy neharmonické [5] kmitočtové charakteristiky hlasového traktu jsou nazývány formanty. Vznikají šířením základního F_0 tónu v hlasovém traktu. Formanty mohou vznikat také rezonancí v dutinách hudebních nástrojů. Naopak oblast lokálního minima v modulu spektra hlásek se nazývá antiformant. Za hlavní formanty jsou považovány první dva, tedy F_1 a F_2 a jsou označovány jako základní. Frekvence základních formantů mají zásadní vliv na rozlišování jednotlivých samohlásek. Mohou být využity v systémech rozpoznávání samohlásek, odhadu věku a pohlaví mluvčího atp.

1.4.3 Lineární prediktivní kódování (LPC)

LPC je jednou z nejefektivnějších metod analýzy řečového signálu, popř. libovolného akustického signálu. Řadí se mezi krátkodobé metody, snaží se tedy odhadnout parametry modelu vytváření řeči z mikrosegmentů původního řečového signálu. Myšlenka LPC je následující: každý n-tý vzorek signálu $s[n]$ lze popsat jako lineární kombinaci Q předchozích vzorků a buzení $u[n]$

$$s[n] = -\sum_{i=1}^{Q} a_i \cdot s[k-n] + G \cdot u[k], \tag{1.4}$$

kde G je koeficient zesílení a Q je řád modelu [10]. Při zpracování LPC jsou nejprve využity psychoakustické aspekty, tedy křivky stejné hlasitosti, maskovací křivky a nelineární vztah mezi intenzitou zvuku a jeho snímanou hlasitostí. LPC koeficienty jsou vypočítání pomocí Levinsonova – Durbinova algoritmu.

1.4.4 Melovské kepstrální koeficienty (MFCC)

Melovské kepstrální koeficienty jsou jedny z nejvýznamnějších řečových příznaků používaných v technikách zpracování řeči. Zpracování MFCC je navrženo tak, aby částečně respektovalo nelineární vlastnosti lidského sluchu. K tomu využívá trojúhelníkovou banku pásmových filtrů s lineárním rozložením frekvencí v tzv. milovské škále. Ta je definována vztahem [10]

$$f_m = 2592 \cdot log_{10}\left(1 + \frac{f}{700}\right), \tag{1.5}$$

kde f [Hz] je frekvence v lineární škále a f_m [mel] je odpovídající frekvence v nelineární milovské škále. Melovské kepstrální koeficienty jsou založeny na principech tvorby řeči v řečovém traktu. MFCC lze určit následujícím způsobem:

- předzpracování signálu zahrnující preemfázi, segmentaci, váhování oknem (nejčastěji Hammingovým)
- vypočtení modulového nebo výkonového spektra
- násobení vypočteného spektra bankou trojúhelníkových filtrů
- zlogaritmování spektra
- vypočet zpětné DCT

1.4.5 Dynamické koeficienty

Všechny příznaky uvedené výše jsou statické, tzn. že příznak je vypočten ze vzorků signálu aktuálně váženého oknem. Ke každému statickému příznaku mohou být určeny ještě tzv. dynamické příznaky, nebo též dynamické koeficienty. Označují se nejčastěji Δc_m a $\Delta^2 c_\mathrm{m}$ a vyjadřují dynamiku časové změny vektorů příznaků [10]. Pro každý

analyzovaný segment se určují lineární regresí z několika po sobě jdoucích segmentů. Většinou se jedná o 3 segmenty. Dynamické koeficienty doplňují statické příznaky, jako MFCC nebo LPC a jsou hojně využívané pro svoji vysokou informační hodnotu a nekorelovatelnost v systémech rozpoznávání řeči.

2 DETEKTORY ŘEČOVÉ AKTIVITY

Detektor řečové aktivity je automatický klasifikátor, který zanalyzuje vstupní řečový signál rozdělený na segmenty a určí, které segmenty obsahují řeč a které nikoli. Tomuto procesu se říká detekce řeč / pauza. Následující kapitola se zabývá několika druhy VAD a popisu jejich funkce.

2.1 Princip detektorů řečové aktivy

Detektory řečové aktivity jsou založené na jednoduchém principu, čítajícím několik kroků (bereme v potaz, že vstupní signál je již předzpracován do požadovaného tvaru):

- načtení segmentů analyzovaného signálu,
- nastavení parametrů nutných k analýze (překrytí, počet inicializačních rámců, atd.)
- vypočtení charakteristiky signálu, závislé na zvoleném typu VAD,
- vypočtení prahové hodnoty z několika počátečních segmentů, neobsahujících řeč (v případě použití adaptivního nastavení prahu),
- porovnání segmentů charakteristiky signálu s prahovou hodnotu,
- je-li prahová hodnota nižší, než charakteristika signálu, daný segment obsahuje řeč; pokud je prahová hodnota vyšší, je detekována tichá pauza a dojde k přepočítání prahové hodnoty (v případě adaptivního prahování).

2.1.1 Adaptivně nastavovaná prahová hodnota

Pokud chceme docílit toho, aby byl řečový detektor plně automatizovaný, musíme prahovou hodnotu vypočítat tak, aby byla přizpůsobená pro daný zkoušený signál. Nejprve se vybere několik počátečních rámců, ve kterých nesmí být řeč. Z těchto rámců se vypočítá střední hodnota $E(\mathrm{X})$ a rozptyl $\sigma^2(\mathrm{X})$. Předpokládejme, že má náhodná veličina X diskrétní rozdělení, kde $\mathrm{P}[\mathrm{X} = \mathrm{x}_i] = \mathrm{p}_i$ a $i \in I$. Potom

$$E(X) = \sum_{i=I} x_i \cdot p_i \quad , \tag{2.1}$$

$$\sigma^2(X) = \sum_{i=I} (x_i - E(X)]^2 \cdot p_i \quad . \tag{2.2}$$

Počáteční prahová hodnota t se získá následovně:

$$t = E(X) + \alpha \cdot \sqrt{\sigma^2(X)}, \tag{2.3}$$

kde $\alpha = \langle 1;2 \rangle$ značí konstantu a její hodnota záleží na druhu detektoru a velikosti SNR.

2.2 Požadavky kladené na detektory řečové aktivity

Systémy pro odstranění šumového signálu vycházejícího z detekce řeči/tiché pauzy jsou zcela závislé na přesném stanovení těchto parametrů. Při špatném určení by mohlo dojít dokonce k zcela opačnému efektu, kdy bychom vstupní signál úplně znehodnotili. Proto by měly detektory řečové aktivity splňovat následující požadavky [7]:

- Detektor by měl být realizovaný ve frekvenční oblasti.
- Detektor by měl umět pracovat v reálném čase.
- Detektor by měl správně pracovat i při malém poměru signál/šum.
- Detektor by měl být výpočetně co nejméně náročný.

2.3 Detektory řečové aktivity pracující v časové oblasti

2.3.1 Energetický detektor

Energetický detektor je nejjednodušší typ VAD. Princip je založen na výpočtu krátkodobé energie jednotlivých segmentů a porovnání s prahovou hodnotou. Mezi jeho nesporné výhody patří snadná implementace, nízká hardwarová náročnost, rychlost detekce a při větším odstupu signálu od šumu i relativně přesná detekce. Nevýhodou energetického detektoru je přílišná citlivost na skokové změny úrovně signálu. Další nevýhoda, která postihuje téměř všechny detektory pracující v časové oblasti, je nepřesná detekce problémových slov. Energie analyzovaných rámců klesnou blízko k průměrné hodnotě šumu, ale pořád se jedná o řečový rámec. Energetický detektor může být využit jako předznačkovač při ručním značení nahrávek (více v kapitole 2.5). K výpočtu se používá vztah

$$E_n = \sum_{k=-\infty}^{\infty} (s[k] \cdot w[n-k])^2, \tag{2.4}$$

kde $s[k]$ je vstupní diskrétní signál, k je pořadí segmentu, nabývajících hodnot $k = 1,2,...,\mathrm{N}$ a $w[n]$ reprezentuje okénkovou funkci. Hodnota E je porovnávána s prahovou hodnotou E_t definovanou vztahem [9]

$$E_t = 1{,}5 \cdot E_b, \tag{2.5}$$

kde E_b je energie pozadí šumu obnovovaná podle vztahu

$$E_b(i+1) = p \cdot E + (1-p) \cdot E_b(i) \tag{2.6}$$

a i je index poslední vzaté hodnoty.

Jestliže $E_n > E_t$, pak n-tý segment obsahuje řečový signál.

Nevýhodou funkce je její velká citlivost na skokové změny úrovně signálu, kdy pauzy vyhodnocuje jako řeč. Tato nepřesnost je způsobena vlivem druhé mocniny na vysokou dynamiku řečového signálu. Tuto nevýhodu eliminuje detektor krátkodobé intenzity, vycházející z detektoru krátkodobé energie. Je popsán vztahem

$$M = \sum_{k=-\infty}^{\infty} |s[k]| \cdot w[n-k] \quad . \tag{2.7}$$

Dle [10] je doporučená délka segmentu 10 až 20 ms a měla by být shodná s délkou okna.

2.3.2 **Detektor krátkodobé funkce středního počtu průchodů nulou**

Detektor krátkodobé funkce středního počtu průchodů nulou spadá do stejné kategorie jako energetický detektor a sdílí jeho výhody i nevýhody.

Krátkodobá funkce středního počtu průchodů nulou je definována jako vážený průměr počtu znaménkových změn řečového signálu uvnitř časového okna. Funkci lze definovat jako

$$Z_n = \sum 0{,}5 |\operatorname{sgn}\{x[m]\} - \operatorname{sgn}\{x[m-1]\}| \cdot w(n-k), \tag{2.8}$$

$$\text{kde } \operatorname{sgn}\{x\} = \begin{cases} 1 & x \geq 0 \\ -1 & x \leq 0 \end{cases}.$$

Z povahy signálu řeči vyplývá fakt, že signál šumu bude mít více průchodů nulovou úrovní, než signál řečový.

Počáteční prahová hodnota se určí ze vztahu 0 nahoře.

2.3.3 **Jednokrokový kepstrální integrální detektor**

Kepstrální detektor je zvláštní případ detektorů pracujících v časové oblasti. Ten totiž využívá Fourierovu transformaci k převedení do frekvenční oblasti a následně zpětnou Fourierovu transformaci pro převod do časové oblasti. V angličtině se tomuto jevu říká quefrency. Pro určení úseku řeč/pauza se porovnává kepstrální vzdálenost $\Delta c[n]$ kepstrálních vektorů $c[n]$ jednotlivých segmentů a obnovované prahové hodnoty $\Delta c_{th}[n]$. Vektory se získají z reálného kepstra. Výhodou kepstrálního detektoru je vysoká schopnost správné detekce i u signálů vysoce kontaminovaných šumem. Nevýhodou může být pomalejší zpracování signálu, protože je nutno provést FFT, logaritmování a zpětnou FFT. Oproti výše zmíněným detektorům lze kepstrální detektor využít pro reálné určení šumu / pauzy, se kterým se může pracovat dále. Např. využít metod odečítání šumu.

Diskrétní spektrum analyzovaného signálu může být určeno diskrétní Fourierovou transformací DFT jako

$$S[k] = DFT\{s[n]\} = \sum_{n=0}^{N-1} s[n] \cdot e^{-j\frac{2\pi}{N}k \cdot n}. \tag{2.9}$$

Dle [14] se pro výpočet spektra signálu nejčastěji používá algoritmu rychlé Fourierovy transformace FFT. Proměnná k značí frekvenční index, nabývajících diskrétních hodnot 0, 1, …, N. N je počet bodů algoritmu Fourierovy transformace. Frekvence každého indexu se vypočítá vztahem $f = (k/N) \cdot f_{vz}$. Reálná složka inverzní diskrétní Fourierovy transformace logaritmu modulu spektrální funkce $S[k]$ se nazývá reálné kepstrum, definované vztahem [9]

$$c[n] = \mathrm{Re}\left\{DFT^{-1}\left[\ln|S[k]|\right]\right\} = \mathrm{Re}\left\{\frac{1}{N}\sum_{k=0}^{N-1} \ln|S[k]| \cdot e^{j\frac{2\pi}{N}k \cdot n}\right\}. \tag{2.10}$$

Kepstrální vzdálenost vektorů se vypočítá jako

$$\Delta c[n] = 4{,}3429 \cdot \sqrt{(c_1[0] - \bar{c}\ [0])^2 + 2 \cdot \sum_{k=1}^{p} (c_1[k] - \bar{c}[k])^2}, \tag{2.11}$$

kde $\bar{c}_k$ je krátkodobý průměr pozadí kepstrálního vektoru.

Rozhodovací úroveň je součtem střední hodnoty kepstrálních vektorů pauzy a rozptylu, tedy:

$$c_{th} = \overline{\Delta c_N}[n] + \alpha \cdot dv(\overline{\Delta c_N}[n]), \tag{2.12}$$

$$\overline{\Delta c_N}[n] = E(X = \Delta c[n]), \tag{2.13}$$

$$dv(\overline{\Delta c_N}[n]) = \sqrt{\sigma^2(X = \Delta c[n])}. \tag{2.14}$$

Rozhodovací proces porovnává hodnoty kepstrální vzdálenosti $\Delta c[n]$ a rozhodovací úrovně c_{th}.

$$OUT = \begin{cases} 0 & když \quad \Delta c[n] < c_{th} \\ 1 & když \quad \Delta c[n] > c_{th} \end{cases}.$$

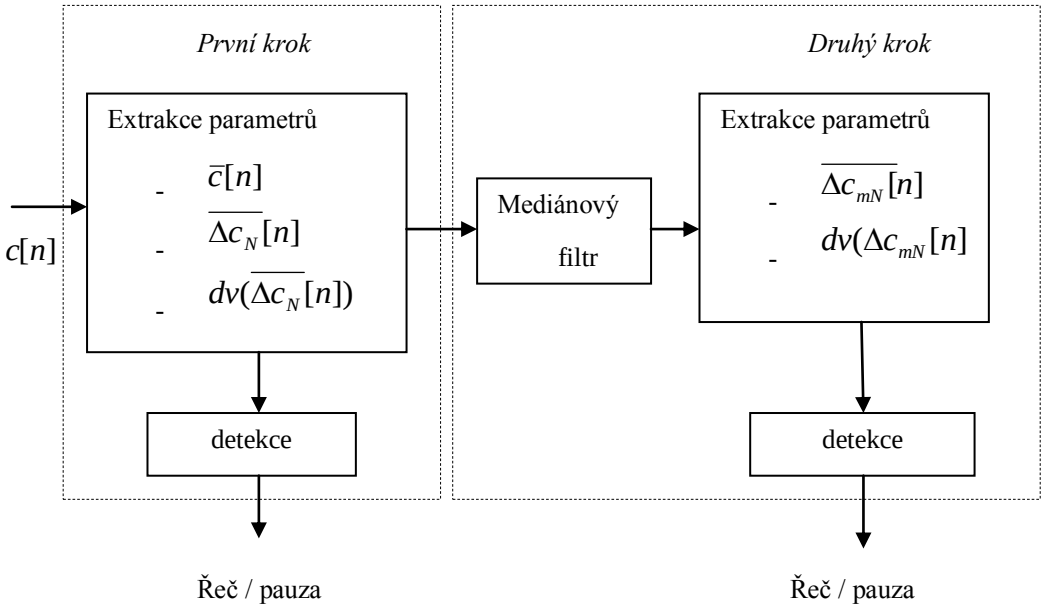
Při detekci pauzy dochází k obnově rozhodovací úrovně

$$\bar{c}[n+1] = (1-\lambda) \cdot \bar{c}[n] + \lambda \cdot \bar{c}[n], \tag{2.15}$$

kde λ představuje konstantu krátkodobého exponenciálního průměrování.

2.3.4 **Dvoukrokový kepstrální integrální detektor**

Motivací pro vytvoření dvoukrokového detektoru byl fakt, že jednokrokový detektor občas detekoval řečový signál jako šum v místech, kde docházelo ke skokovým změnám šumu. Rozdíl mezi kepstrálními detektory je znázorněn na následujícím obrázku. Dvoukrokový algoritmus nejprve extrahuje kepstrální koeficienty ze vstupního signálu a určí kepstrální vzdálenosti, poté je vyhladí mediánovým filtrem a znovu přepočítá rozptyl a střední hodnotu vektorů kepstrálních vzdáleností. Vzorce pro výpočet jsou stejné jako u jednokrokového detektoru.



Obr. 2.1: Blokové schéma kepstrálních detektorů

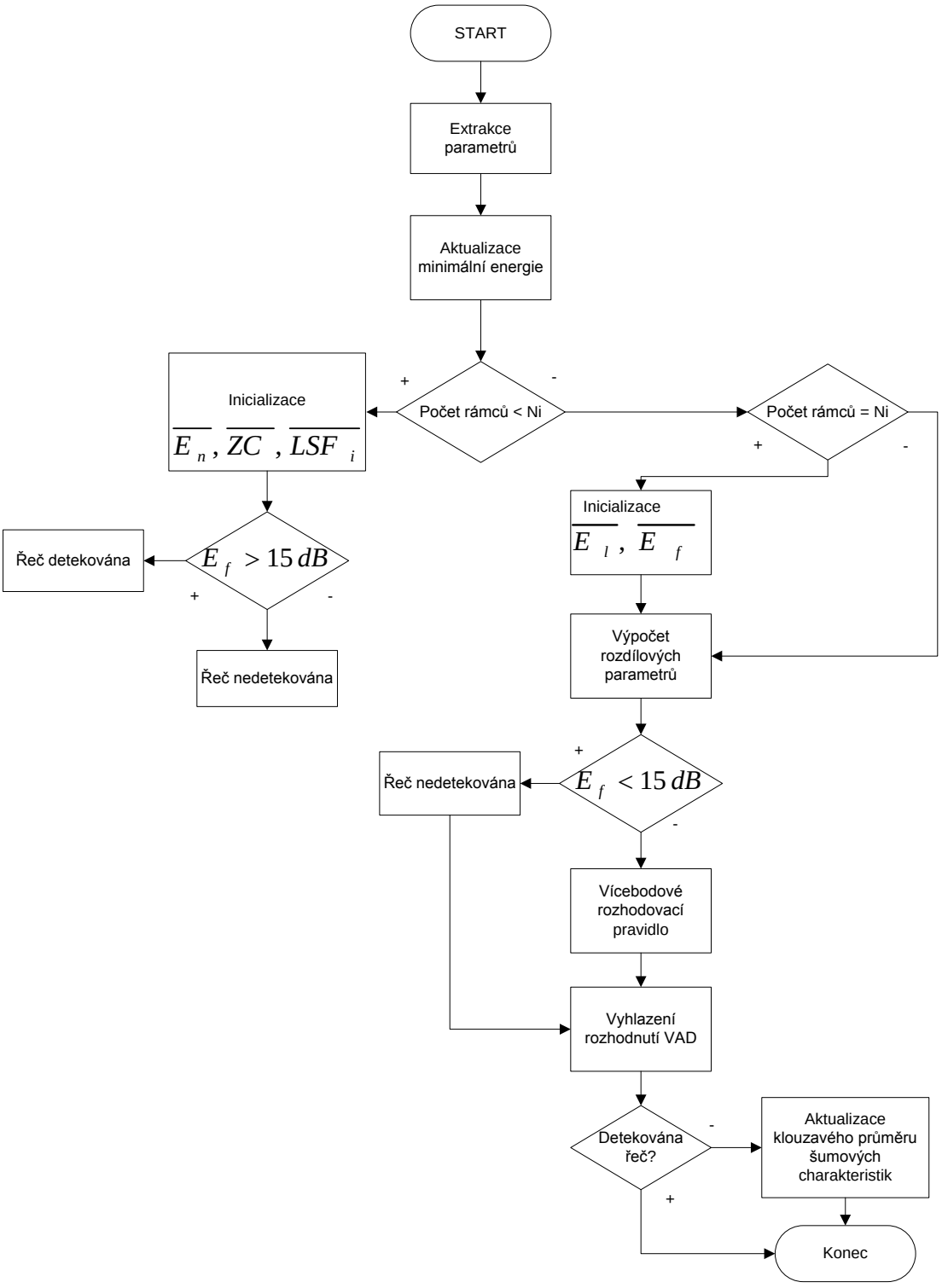
2.4 Detektory řečové aktivity pracující ve frekvenční oblasti

2.4.1 **Detektor ITU-T G.729**

Tato část se věnuje pouze samotné funkci detektoru, ne jeho dalším částem, které se aktivují při detekci šumu. Detektor pracuje následovně: provádí rozhodování každých 10 sekund v rámci o velikosti 240 vzorků. K detekci se extrahuje soubor parametrů, jež jsou použity pro počáteční rozhodování. Tyto parametry jsou: celková energie, úzkopásmová energie, krátkodobé funkce středního počtu průchodů nulovou úrovní a kmitočet spektrálních párů. Při detekci šumového rámce dochází k aktualizaci klouzavého průměru pozadí šumu, aby v budoucnu nedošlo k nesprávné indikaci šumového rámce. Z každého rámce se vždy získají rozdílové parametry. Ty jsou dány

jako rozdíl aktuálně získaných hodnot a jejich průměrných hodnot. Rozhodovací proces rozhoduje na základě vícebodového rozhodovacího pravidla, které postupně testuje všechny parametry. Vyhlazením rozhodnutí řečového detektoru získám finální výsledek. Výstup detektoru ITU-T G.729 je stejně jako ostatní detektory tvořen jedničkami a nulami, symbolizujícími řeč či pauzu. Jak je z popisu zřejmé, je detektor G.729 náročnější na výpočetní výkon a rychlost detekce je nižší než u výše popsaných detektorů. Další zdánlivou nevýhodou může být použití velkého množství parametrů. Při špatném nastavení parametrů dojde k velice nepřesné detekci. Rozdíl v přesnosti detekce mezi správným a nesprávným nastavením může být i 20 %. Výhody tohoto detektoru jsou: pružnost detektoru v různých podmínkách, správná detekce problematických slov, vysoká účinnost detektoru. Detektor G.729 je součástí kodeku pro kompresi digitalizovaného audio signálu ITU-T G.729. Ten se využívá např. ve VoIP telefonii. V současné době existuje několik verzí ITU-T G.729, počínaje verzí Annex A a konče verzí Annex I.

Podrobnému popisu algoritmu se věnuje následující část této kapitoly a jednotlivé kroky jsou zobrazeny ve vývojovém diagramu na Obr. 2.2.



Obr. 2.2: Vývojový diagram rozhodovacího procesu detektoru ITU-T G.729

Předzpracování vstupního signálu

Vstupní signál je nejprve filtrován filtrem typu horní propust, který slouží jako ochrana

proti nechtěným nízkofrekvenčním složkám signálu. Toho je dosaženo filtrem $H_{h1}(z)$ druhé úrovně s mezní frekvencí 140 Hz. Takto upravený signál bude dále značen $s[n]$.

$$H_{h1}(z) = \frac{0{,}46363718 - 0{,}92724705z^{-1} + 0{,}46363718z^{-2}}{1 - 1{,}9059465z^{-1} + 0{,}9114024z^{-2}} \qquad (2.16)$$

```
signal = filter([0.4636718 -0.92724705 0.4636718],[1 -0.9059465
0.9114024],y);
```

Signál $s[n]$ je následně vážen nesymetrickou okenní funkcí

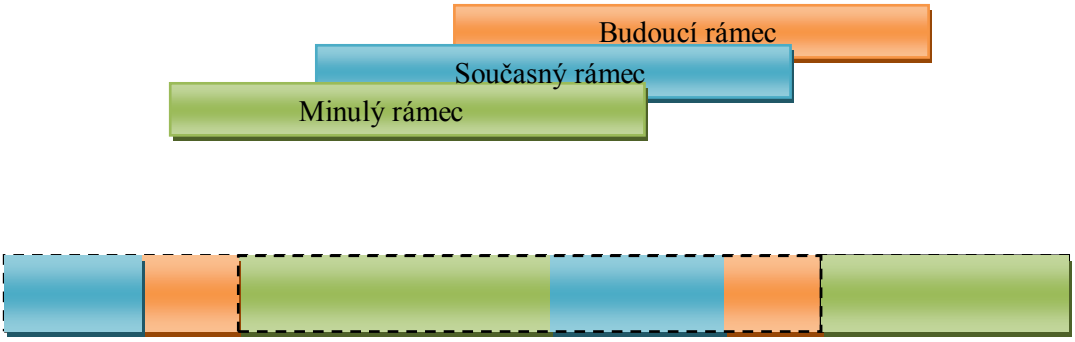
$$w_{lp}[n] = \begin{cases} 0{,}54 - 0{,}46cos\left[\frac{2\pi n}{399}\right] & n = 0, \ldots, 199 \\ \cos\left[\frac{2\pi(n-200)}{159}\right] & n = 200, \ldots, 399 \end{cases},$$

kde horní část představuje Hammingovo okno a spodní část funkce představuje čtvrtinu periody kosinové funkce.

```
wlp1 = 0.54-0.46*cos((2*pi*([1:200]-1))/399);
wlp2 = cos((2*pi*(([201:240]-1)-200))/159));
wlp = [wlp1 wlp2];
```

Extrakce parametrů

Z každého rámce řečového signálu je extrahován soubor parametrů – lineárních predikčních koeficientů – pomocí LP analýzy. Linerání predikční analýza pracuje s oknem tvořeným 120 vzorky minulého rámce, 80 vzorky současného rámce a 40 vzorky budoucího rámce. Přítomnost budoucího rámce o velikosti 40 vzorků vnáší do algoritmu zpoždění přesně 5 ms. Skladba okna je znázorněna na Obr. 2.3



Obr. 2.3: Rámec LP analýzy tvořený okny

Signál vážený oknem:

$$s[n] = w_{lp}[n]s[n] \;\; n = 0, \dots, 239 \tag{2.17}$$

Je použit pro výpočet autokorelačních koeficientů:

$$r(k) = \sum_{n=k}^{239} s'[n]s'[n-k] \;\; k = 0, \dots, 10 \tag{2.18}$$

```
r = autocorr(segment(i,:), 10);
```

Abychom se vyhnuli aritmetickým problémům způsobených signálem o nízké vstupní energii, nastavíme hodnotu prvního autokorelačního koeficientu na jedna:

$$r(0) = 1{,}0.$$

Poté je aplikována metoda rozšíření šířky pásma násobením autokorelačních koeficientů s:

$$w_{lag}(k) = exp\left[-\frac{1}{2}\left(\frac{2\pi f_0 k}{f_s}\right)^2\right] \;\; k = 1, \dots, 10 \tag{2.19}$$

kde f_s značí vzorkovací frekvenci řečového signálu a $f_s = 60Hz$. Dále je první autokorelací koeficient $r(0)$ násoben korekčním faktorem bílého šumu. Modifikované autokorelací koeficienty budou tedy vypadat následovně:

$$r'(0) = 1{,}0001r(0)$$

$$r'(k) = w_{lag}(k)r(k) \;\; k = 1, \dots, 10.$$

```
wlag = exp(-0.5*((2*pi*60*(1:10))/Fs).^2);
r(1) = r(1)*1.0001;
r(2:11) = r(2:11)*.^wlag;
```

Kmitočet spektrálních párů (LSF)

Soubor lineárních predikčních koeficientů je odvozen z autokorelace a z nich je poté odvozena skupina spektrálních párů $\{LSF\}_{i=10}^{p}$, kde p = 10, jak je popsáno v [2]. K výpočtu je využit Levinson-Durvinův algoritmus.

```
LP = levinson(r);
LSF(i) = poly2lsf(LP);
```

Celková energie

Celková energie rámce se vypočítá jako logaritmus prvního autokorelačního koeficientu $r'(0)$, tedy:,

$$E_f = 10\log_{10}\left[\frac{1}{\mathrm{N}}\mathrm{r}'(0)\right], \tag{2.20}$$

kde N = 240 je velikost LPC okna.

```
Ef(i) = 10*log10(r(1)/N);
```

Úzkopásmová energie

Úzkopásmová energie E_l je počítána v pásmu 0 až f_l herztů, podle následujícího vztahu:

$$E_l = 10log_{10}\left[\frac{1}{\mathrm{N}}\mathrm{h}^T R\mathrm{h}\right], \tag{2.21}$$

kde h značí impulzovou odezvu FIR filtru s mezní frekvencí f_l [Hz], R je Toeplitzova autokorelační matice s autokorelačními koeficienty na diagonále.

```
R = toeplizt(r);
El = 10*log10((1/240)*h*R*h');
```

Střední počet průchodů nulovou úrovní

Normalizovaný výpočet středního počtu průchodů nulovou úrovní v rámci je dán následovně:

$$ZC = \frac{1}{2M}\sum_{i=0}^{M-1}[|sgn[x(i)] - sgn[x(i-1)]|], \tag{2.22}$$

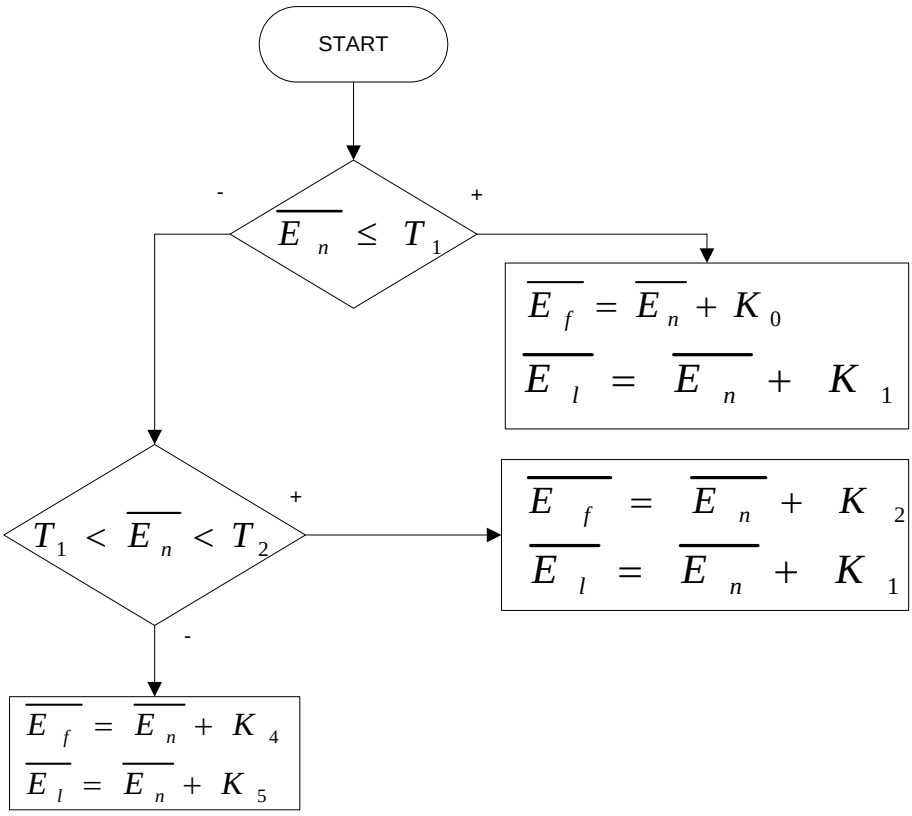
kde $\{x(i)\}$ představuje předzpracovaný vstupní signál a M = 80.

```
pf = segment(i,121:200);  %výběr současného rámce o délce 80
ZC(i) = (sum(abs(diff(pf>0)))/length(pf));
```

Inicializace klouzavého průměru pro šumové charakteristiky pozadí

V prvních Ni rámcích jsou spektrální parametry šumových charakteristik pozadí, značeny jako $\{\overline{LSF_i}\}_{i=1}^{p}$, získány jako průměr $\{LSF_i\}_{i=1}^{p}$ v p=*N*i rámcích. Podobně vypočteme klouzavý průměr středního počtu průchodu nulovou úrovní $\overline{ZC}$ jako průměr ZC v *N*i rámcích. Klouzavý průměr energie šumu pozadí $\overline{E_f}$ a klouzavý průměr úzkopásmové energie šumu pozadí $\overline{E_l}$ jsou získány následovně:

1) Nejprve inicializační procedura použije $\overline{E_n}$, definovanou jako průměrnou hodnotu celkové energie v prvních *Ni* rámcích. Klouzavé průměry $\{\overline{LSF_i}\}_{i=1}^{p}$, $\overline{ZC}$ a $\overline{E_n}$ využívají pouze rámce s celkovou energií větší než 15 dB.
2) Inicializační procedura projde sérií podmínek a upraví hodnoty klouzavého průměru.
3) Rozhodovací proces pro aktualizaci klouzavých průměrů úzkopásmové energie a celkové energie rámce jsou vyobrazeny v následujícím vývojovém diagramu:



Obr. 2.4: Vývojový diagram procesu aktualizace klouzavých průměrů

Přehled hodnot všech konstant K_x, T_x, a, b se nachází v příloze dokumentu.

Výpočet dlouhodobé minimální energie

Dlouhodobá minimální energie $E_{\min}$ je vypočítána jako minimum celkové energie E_f z N_0 předcházejících rámců. Protože je hodnota N_0 relativně velké číslo, $E_{\min}$ získává hodnoty z minima celkové energie E_f z předchozích výpočtů.

```
Emin = min(Ef(i-N0:i));
```

Výpočet rozdílových parametrů

Rozdílové parametry jsou vypočítány z parametrů současných rámců a klouzavého průměru šumu pozadí daného parametru.

Spektrální rozložení ΔS

Spektrální rozložení je dáno jako suma druhé mocniny z rozdílu $\{LSF_i\}_{i=1}^{p}$ a $\{\overline{LSF_i}\}_{i=1}^{p}$ současného rámce:

$$\Delta S = \sum_{i=1}^{p} (LSF_i - \overline{LSF_i})^2 \qquad (2.23)$$

Rozdíl celkové energie ΔE_{f}

Rozdílová hodnota celkové energie je počítán jako rozdíl celkové energie E_{f} a klouzavého průměru energie šumu pozadí $\overline{E}_{\mathrm{f}}$ aktuálního rámce:

$$\Delta E_{\mathrm{f}} = \overline{E}_{\mathrm{f}} - E_{\mathrm{f}} \tag{2.24}$$

Rozdíl úzkopásmové energie ΔE_{l}

Rozdílová hodnota úzkopásmové energie je počítán jako rozdíl úzkopásmové energie E_l a klouzavého průměru úzkopásmové šumu pozadí $\overline{E_l}$ aktuálního rámce:

$$\Delta E_{\mathrm{l}} = \overline{E}_{\mathrm{l}} - E_{\mathrm{l}} \tag{2.25}$$

Rozdíl středního počtu průchodů nulovou úrovní ΔZC

Rozdílová hodnota středního počtu průchodů nulovou úrovní je počítán jako rozdíl ZC a $\overline{ZC}$ aktuálního rámce:

$$\Delta ZC = \overline{ZC} - ZC \tag{2.26}$$

Vícebodové rozhodovací pravidlo

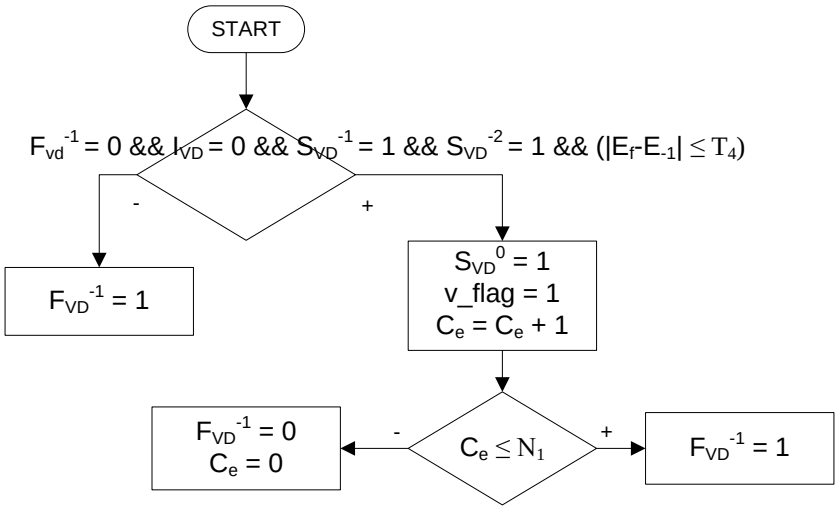
Vícebodové rozhodovací pravidlo se značí I_{VD} a nabývá hodnot 0 a 1. Nula značí nepravdu, tedy rámec neobsahuje řeč a jedna značí pravdu, tedy rámec obsahuje řeč. Počáteční hodnota $I_{\mathrm{VD}} = 0$. Vícebodové rozhodovací pravidlo postupně prochází 14 podmínek. Pokud je podmínka pravdivá, I_{VD} se nastaví na jedna. V opačném případě se testuje další podmínka. Jestliže ani jedna podmínka není splněna, I_{VD} si ponechá hodnotu nula.

1. Pokud $\Delta S > \mathrm{a}_1 \cdot \Delta ZC + \mathrm{b}_1$, potom $\mathrm{I_{VD}} = 1$
2. Pokud $\Delta S > \mathrm{a}_2 \cdot \Delta ZC + \mathrm{b}_2$, potom $\mathrm{I_{VD}} = 1$
3. Pokud $\Delta E_{\mathrm{f}} < \mathrm{a}_3 \cdot \Delta ZC + \mathrm{b}_3$, potom $\mathrm{I_{VD}} = 1$
4. Pokud $\Delta E_{\mathrm{f}} < \mathrm{a}_4 \cdot \Delta ZC + \mathrm{b}_4$, potom $\mathrm{I_{VD}} = 1$
5. Pokud $\Delta E_{\mathrm{f}} < \mathrm{b}_5$, potom $\mathrm{I_{VD}} = 1$
6. Pokud $\Delta E_{\mathrm{f}} < \mathrm{a}_6 \cdot \Delta S + \mathrm{b}_6$, potom $\mathrm{I_{VD}} = 1$
7. Pokud $\Delta S > \mathrm{b}_7$, potom $\mathrm{I_{VD}} = 1$
8. Pokud $\Delta E_{\mathrm{l}} < \mathrm{a}_8 \cdot \Delta ZC + \mathrm{b}_8$, potom $\mathrm{I_{VD}} = 1$
9. Pokud $\Delta E_{\mathrm{l}} < \mathrm{a}_9 \cdot \Delta ZC + \mathrm{b}_9$, potom $\mathrm{I_{VD}} = 1$
10. Pokud $\Delta E_{\mathrm{l}} < \mathrm{b}_{10}$, potom $\mathrm{I_{VD}} = 1$
11. Pokud $\Delta E_{\mathrm{l}} < \mathrm{a}_{11} \cdot \Delta S + \mathrm{b}_{11}$, potom $\mathrm{I_{VD}} = 1$
12. Pokud $\Delta E_{\mathrm{l}} > \mathrm{a}_{12} \cdot \Delta E_{\mathrm{f}} + \mathrm{b}_{12}$, potom $\mathrm{I_{VD}} = 1$
13. Pokud $\Delta E_{\mathrm{l}} < \mathrm{a}_{13} \cdot \Delta E_{\mathrm{f}} + \mathrm{b}_{13}$, potom $\mathrm{I_{VD}} = 1$
14. Pokud $\Delta E_{\mathrm{l}} < \mathrm{a}_{14} \cdot \Delta E_{\mathrm{f}} + \mathrm{b}_{14}$, potom $\mathrm{I_{VD}} = 1$

Vyhlazení výstupu detektoru

Aby výsledek odrážel dlouhodobou stacionární povahu řečového signálu, je rozhodnutí detektoru vyhlazeno. Indikátor rozhodnutí v_flag hlídá, zda se předchozí a současné rozhodnutí liší, nebo ne. Vyhlazené rozhodnutí detektoru současného rámce, předcházejícího rámce a rámce předminulého se označuje jako S_{vd}^{0}, S_{vd}^{-1} a S_{vd}^{-2}. Proměnné S_{vd}^{-1} a S_{vd}^{-2} mají hodnotu 1. S_{vd}^{0} je rovno vyhodnocení VAD. Proces vyhlazení sestává ze 4 kroků:

1) Pokud $(I_{\mathrm{VD}} = 0)$ a $\left(S_{\mathrm{vd}}^{-1} = 1\right)$ a zároveň $(E > \overline{E}_{\mathrm{f}} + \mathrm{T}_3)$, potom
   $S_{\mathrm{VD}}^{0} = 1$
   $v_{\mathrm{flag}} = 1$
2) Ve druhém kroku je zavedena nová logická proměnná F_{vd}^{-1} a proměnná zaznamenávající počet vyhlazení C_{e}. Počáteční nastavení $C_{\mathrm{e}} = 1$. Dále je vytvořena proměnná s hodnotou celkové energie předcházejícího rámce E_{-1}. Druhý krok sestává z následujících podmínek:



Obr. 2.5: Vývojový diagram procesu vyhlazení detekce řečové aktivity - druhý krok

3) Ve třetím kroku je zavedeno počítadlo pokračování šumu C_{s}. Na začátku je nastaveno na nulu, a pokud je současný rámec označen detektorem jako šum, C_{s} se zvýší o jedna.
   Pokud $((S_{\mathrm{VD}}^{0} = 1)$ a $(C_{\mathrm{s}} > \mathrm{N}_2)$ a zároveň $(E_{\mathrm{f}} - E_{-1} \leq \mathrm{T}_5))$, potom
   $S_{\mathrm{VD}}^{0} = 0$
   $C_{\mathrm{s}} = 0$
   Pokud $(S_{\mathrm{VD}}^{0} = 1)$, potom $C_{\mathrm{s}} = 0$
4) Pokud $((\mathrm{I_f} < \overline{E}_{\mathrm{f}} + \mathrm{T}_6)$ a $(frm\_count > \mathrm{N}_0)$ a $(v\_flag = 0))$, pak $S_{\mathrm{VD}}^{0} = 0$

Aktualizace klouzavého průměru šumové charakteristiky pozadí

Aktualizace klouzavého průměru se provádí jako poslední krok detektoru. Testuje se následující podmínka:

Pokud ($E_f < \overline{E_f} + T_6$), potom proveď aktualizaci.

Při aktualizaci se využívá autoregresivních koeficientů prvního řádu. Pro odlišné parametry se používají odlišné autoregresivní koeficienty a podle povahy testovaného rámce v závislosti na předchozí detekci se jednotlivé koeficienty ještě mění. Nechť β_{Ef} je AR koeficient pro aktualizaci $\overline{E_f}$, β_{El} je AR koeficient pro aktualizaci $\overline{E_l}$, β_{ZC} je AR koeficient pro aktualizaci $\overline{ZC}$ a nakonec β_{LSF} je AR koeficient $\{LSF\}_{i=1}^{p}$. Celkový počet rámců, ve kterých dojde k vyhlazení, je zaznamenáván do proměnné C_n. Na základě hodnoty proměnné C_n se mění parametry AR koeficientů. Aktualizace klouzavého průměru šumových charakteristik pozadí je popsána následujícími vztahy:

$$\overline{E_f} = \beta_{Ef} \cdot \overline{E_f} + (1 - \beta_{Ef}) \cdot E_f$$

$$\overline{E_l} = \beta_{El} \cdot \overline{E_l} + (1 - \beta_{El}) \cdot E_l$$

$$\overline{ZC} = \beta_{ZC} \cdot \overline{ZC} + (1 - \beta_{ZC}) \cdot ZC$$

$$\overline{LSF_i} = \beta_{LSF} \cdot \overline{LSF_i} + (1 - \beta_{LSF}) \cdot LSF_i,$$

kde $i = 0, 1, 2 \ldots p$. Hodnoty $\overline{E_f}$ a C_n jsou dále aktualizovány dle podmínky:

pokud (současné pořadové číslo rámce > N_0) a zároveň ($\overline{E_f} < E_{min}$) , potom

$\overline{E_f} = E_{min}$

$C_n = 0$.

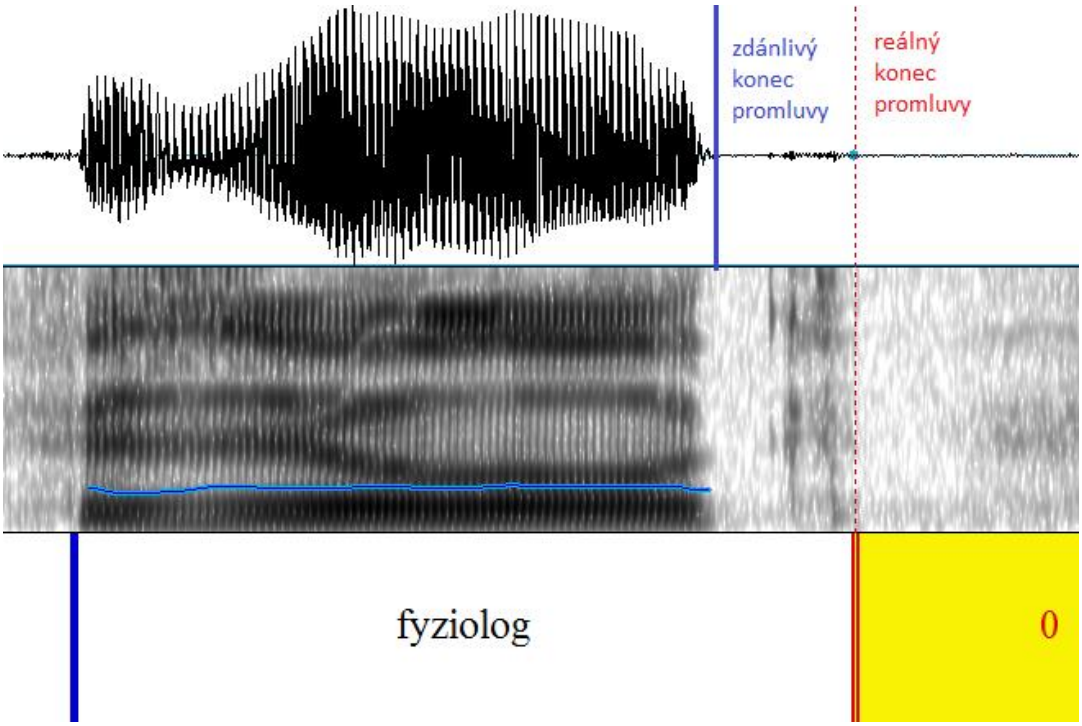
2.5 Ruční vymezení řeči

Ručním vymezení řeči se myslí způsob ručního stanovení hraničních bodů promluvy, nejčastěji pomocí popisových os. Důležitost této úlohy tkví zejména v rozpoznávání krátkých izolovaných slov a krátkých frází. Svoje uplatnění může nalézt i při značení souvislé řeči. Takto označené nahrávky mohou být totiž využity jako trénovací množina pro statistické metody detekce řečové aktivity. Statistickými metodami se myslí takové metody, kdy jsou prahové hodnoty nastaveny na základě trénovacích dat. Existuje několik jevů, které komplikují počítačem prováděné detekování hraničních bodů promluvy [10]:

- Akustické prostředí s větší mírou šumu na pozadí (vzdálený hovor kolemjdoucích, hluk projíždějícího automobilu atp.)
- Rušení způsobené nedokonalostí přenosového kanálu (tónové interference)
- Nedokonalá artikulace, vyvolaná mluvčím, způsobuje tzv. doprovodné zvuky

(hlasité nadechování, mlaskání atp.), jejichž energie může být podobná jako energie samotné promluvy

Zvláštním případem jsou slova, jejichž začátek a konec je obtížně definovatelný. V takovém případě má začátek a konec slova velice malou energii a detektory řečové aktivity založené na výpočtu prahové hodnoty z energie rámce by tyto části označily jako tichou pauzu. Podobně se může zmýlit i člověk, provádějící ruční vymezování řeči. Řešením je využití spektrálních charakteristik signálu, kdy jsou vidět spektrální změny signálu. Problém ručního značení je naznačeno na Obr. 2.6 na příkladu slova fyziolog, provedeno v programu Praat.



Obr. 2.6: Ukázka ručního značení začátku a konce promluvy v programu Praat

3 METODY ZVÝRAZNĚNÍ ŘEČI

V této fázi již víme, které úseky signálu obsahují řeč, a které šumový signál. Než je ale možné pustit se do separace řeči ze šumového prostředí, je potřeba si tyto metody rozdělit do dvou skupin:

- jednokanálové metody
- vícekanálové metody

Jednokanálové metody pracují s řečovým signálem kontaminovaným šumem, který je vzat pouze v jednom místě prostoru. Tedy takový signál, který byl zaznamenán pouze jedním mikrofonem. Celkový šum, aditivně nebo konvolučně přičtený k signálu, je ve výsledku směsí různých šumů, hluků, rušení a odrazů. Tento fakt znesnadňuje, až téměř znemožňuje nalezení optimální metody potlačení hluku fungující pro všechny typy rušení. Po potlačení šumu některou z metod zůstává v signálu hudební šum. Hudební šum se projevuje nestacionárními harmonickými tóny, které připomínají cvrlikání ptáků [7]. Mezi jednokanálové metody patří:

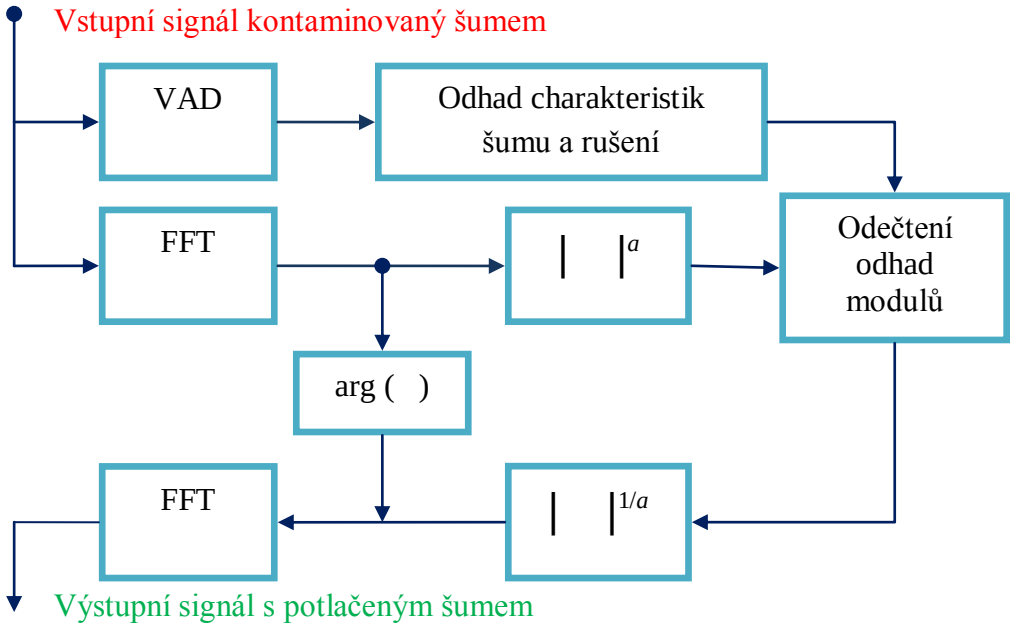
- metoda spektrálního odečítání
- metoda RASTA (RelAtive SpecTrAl)
- metoda mapování spektrogramu
- číslicová filtrace

Vícekanálové metody zaznamenávají signál z většího počtu mikrofonů (dvou a více) rozmístěných v oblasti, kde chceme daný signál zachytit. Říkáme jim také senzorová pole. Díky znalosti prostorového umístění daného rušivého signálu ho můžeme potlačit pomocí směrové adaptace. Výhoda vícekanálové metody je jasná, a to je dosažení mnohem přesnějších výsledků odstranění šumu než u jednokanálové metody. Na druhou stranu se zvyšuje výpočetní náročnost celého procesu, nutnost použití drahé a komplikované soustavy mikrofonů, složitá instalace celého pole, dodržení prostorové vzorkovací poučky apod. Mezi vícekanálové metody patří:

- analýza nezávislých komponent (ICA)
- analýza řídkých komponent (SCA)
- metoda tvarování přijímací charakteristiky (Beam-Forming)

3.1 Metoda spektrálního odečítání

Tato metoda patří k oblíbeným a často používaným metodám díky své nízké výpočetní náročnosti a pružnosti v reálných podmínkách. Výsledek závisí na vhodně zvolených parametrech a na tom, zda spolu šum a řeč korelují, či nikoliv. Obr. 3.1 znázorňuje blokové schéma základních funkcí spektrálního odečítání.



Obr. 3.1: Blokové schéma metody spektrálního odečítání

Postup při zpracování řeči metodou spektrálního odečítání je následující. Vstupní diskrétní signál je rozdělen pomocí časového okna $w[n]$ na segmenty a ty jsou postupně zpracovávány. Okno se volí buď Hammingovo, nebo Hannovo. Bývá vhodné zvolit překrytí segmentů, nejčastěji 40 až 50 % svojí délky. Vypočítá se obraz DFT pomocí FFT každého segmentu podle

$$X_i[k] = \sum_{n=0}^{N-1} w[n] \cdot x_i[n] \cdot e^{-j\frac{2\pi}{N}n}, \tag{3.1}$$

kde $k = 0,1,...,N, N-1$. Parametr i určuje počet rámců výchozího řečového signálu $x[n]$, který je zpracováván. Tento proces je označen jako krátkodobá spektrální analýza, Spektrum jednoho rámce lze zapsat jako

$$X_i[k] = \left|X_i[k]\right| \cdot e^{j \cdot \arg\{x_i[k]\}}, \tag{3.2}$$

kde $\left|X_i[k]\right|$ je modul spektra i-tého rámce a $\arg\{X_i[k]\}$ je jeho argument [12]. Argument je ponechán beze změny a při rekonstrukci signálu se přičítá zpět. Je zpracováván pouze modul signálu $\left|X_i[k]\right|$. Ze segmentů zpracovávaného signálu, které detektor řečové aktivity označí jako šum, je průběžně oceňováno a obnovováno spektrum šumu. Provedeme výkonové spektrální odečítání, pro něž platí

$$\left|Y_i[k]\right|^a = \left|X_i[k]\right|^a - E\left\{\left|N_i[k]\right|^a\right\}, \tag{3.3}$$

kde a je parametr, volený libovolně. Nejčastěji se používá $a = 2$ pro výkonové spektrální odečítání nebo $a = 1$ pro amplitudové spektrální odečítání. Volba parametru má výrazný vliv na výsledné potlačení šumu. Řekněme, že $a = 2$. $|X_i[k]|^a$ bude značit výkonové spektrum řeči kontaminované aditivním šumem a $\mathrm{E}\{|N_i[k]|^a\}$ střední hodnotu výkonového spektra aditivního šumu. Alternativně je možné $\mathrm{E}\{|N_i[k]|^a\}$ získat jako výstup rekurzivního filtru dolní propusti prvního řádu

$$\mathrm{E}\{|N_i[k]|^a\} = \mathrm{E}\{|N_i[k,t]|^a\} = \xi \cdot \mathrm{E}\{|N_i[k,t-1]|^a\} + (1-\xi) \cdot \mathrm{E}\{|N_i[k,t]|^a\} \quad (3.4)$$

kde t značí diskrétní čas a $\xi \in \langle 0{,}8; 0{,}99 \rangle$.

Při malém poměru signálu ku šumu se může objevit slabina metody spektrálního odečítání, tedy veliká citlivost na změny ve spektru šumu. Projevuje se negativním odhadem amplitudového nebo výkonového spektra $|Y_i[k]|^a$. Nastane tak v případě velké změny ve spektru šumu a střední hodnota výkonového spektra šumu bude větší než výkonové spektrum řeči a šumu. Negativní odhady spektra řeči odstraňuje mapovací funkce, která vypadá následovně:

$$|Y_i[k]|^a = \begin{cases} |X_i[k]|^a - \alpha \cdot |E\{N_i[k]\}|^a, & \text{jestliže } |Y_i[k]|^a > \beta \cdot |X_i[k]|^a \\ \beta \cdot |X_i[k]|^a & \text{v ostatních případech,} \end{cases}$$

kde β je vhodně zvolený parametr (volí se obvykle v hodnotách setin) [odkaz Psutka] a α je parametr nabývající hodnot $\langle 1; 3 \rangle$. Poté je provedena zpětná úprava odmocněním modulu $a\sqrt{|Y_i[k]|}$ parametrem a k získání modulu spektra extrahované řeči $|Y_i[k]|$. K odmocněnému modulu přidáme nezměněný argument $\arg\{X_i[k]\}$ a provedeme zpětnou Fourierovu transformaci všech rámců pomocí vztahu

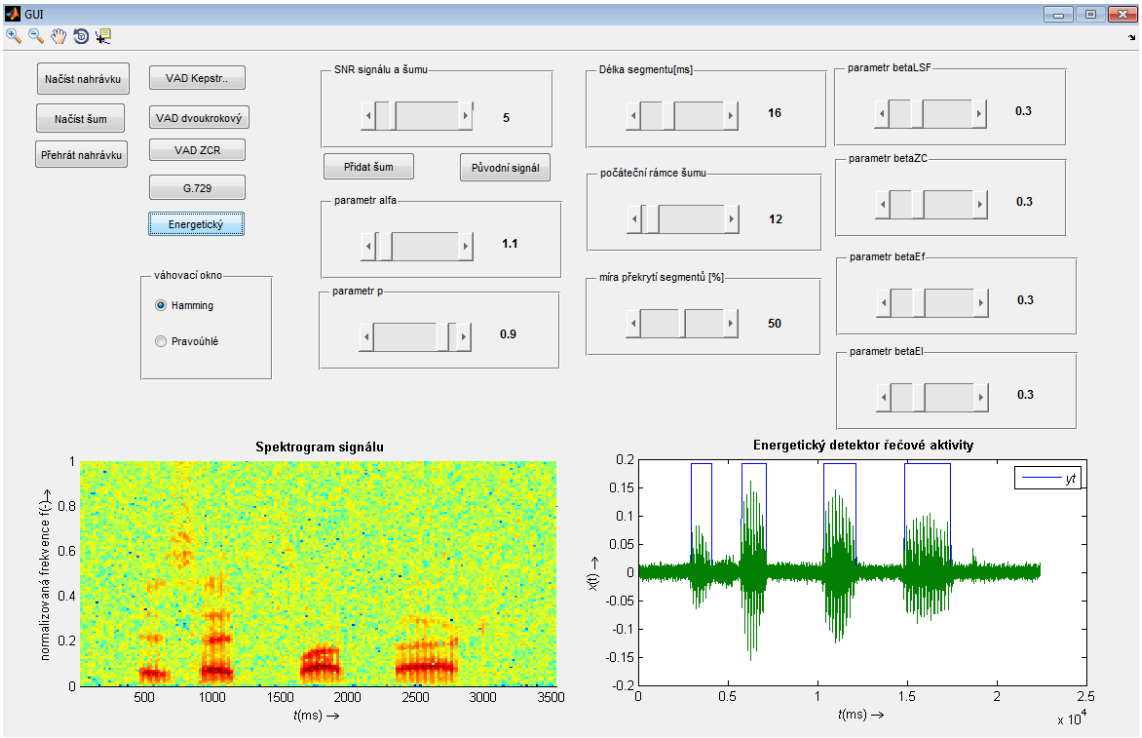
$$s_i[n] = \frac{1}{N} \sum_{k=0}^{N-1} |Y_i[k]| \cdot e^{j \cdot \arg\{x_i[k]\}} \cdot e^{j \cdot k \frac{2\pi}{N} n}, \quad (3.5)$$

$k = 0,1,2,\ldots N, N-1.$

Tím dojde ke spojení segmentů do původního celku. Při špatně zvolených parametrech může dojít k degradaci signálu, protože se při spektrálním odečítání v upraveném signálu generuje v menší či větší míře tzv. hudební šum. Ten se projevuje kovovým zvukem. Je to proto, že odhady parametrů filtru spektrálního odečítání mají odlišnou přesnost pro různá frekvenční pásma signálu [10]. Existuje několik možností, jak hudební šum snížit. Nejjednodušším způsobem je snížit rozptyl odhadu za pomocí dílčího průměrování modulu spektra zvýrazněné. Dalším způsobem může být použití nelineární metody spektrálního odečítání, kdy je odhad výkonové spektrální hustoty násoben kmitočtově závislou konstantou. Nastavením maskovacích vlastností díky šumovému maskovacímu prahu posluchač hudební šum pod tímto prahem neslyší [12].

4 PRAKTICKÁ ČÁST

Metody zvýraznění řeči a odfiltrování šumu jsou závislé na kvalitě rozpoznávání úseků řeč / pauza. Bylo proto vytvořeno několik detektorů řečové aktivity v programu Matlab (viz Obr. 4.1) a provedena série testů. Testy byly prováděny s dvoukrokovým kepstrálním detektorem a detektorem ITU-T G.729. K tomu byly využity dvě sady nahrávek. První sada je databáze TIMIT a druhá byla vytvořena studentem VUT, Bc. Pavlem Pelikánem, v rámci jeho bakalářské práce (dále jako databáze VUT). Všechny nahrávky obsahují téměř čistý řečový signál s minimem šumu. Dále budou označovány jako referenční. Ke každé testované entitě bylo aditivně přičteno několik úrovní šumu, o různých SNR. Detektory byly testovány pro tyto různé úrovně šumu. Výsledkem tohoto testu bylo zjistit optimální nastavení parametrů detektorů. Porovnání vlivu nastavení detektoru bylo provedeno pomocí metody spektrálního odečítání zašuměné nahrávky.



Obr. 4.1: Grafické rozhraní programu pro detekci úseků šum/pauza

4.1 Testovací metodika

Testování detektorů řečové aktivity a hledání optimálních parametrů se skládá z těchto kroků:

1) Výběr testovací databáze
2) Přičtení k testovacím vzorkům vhodné typy rušení
3) Aplikování detektoru řečové aktivity na připravené vzorky s postupným nastavováním parametrů a zaznamenáváním nejlepších výsledků
4) Stanovení korekční hranice a filtrace výsledků převyšující tuto hranici
5) Aplikace metody spektrálního odečítání na náhodné vzorky

V této práci byly použity dvě testovací databáze – TIMIT a databáze vzniklá na půdě VUT. Obě databáze splňují následující podmínky:

- Velký počet nahrávek
- Dostatečná kvalita nahrávek
- Nahrávky musí být namluveny oběma pohlavími
- Úvodní část nahrávky musí být tvořena pouze šumem, aby mohlo dojít ke správné inicializaci parametrů VAD

O testovacích databázích pojednávají následující kapitoly. V tuto chvíli je k dispozici vhodná databáze nahrávek. Je potřeba k nahrávkám aditivně přičíst různé druhy rušení. V této práci jsou použity tři druhy rušení:

- bílý šum
- hluk mixéru (úzkopásmový šum)
- zvuk zapnuté sprchy (širokopásmový šum)

Přičtení šumu k nahrávce bylo realizováno v prostřední Matlab funkcí addnoise

```
[noisy, noise] = addnoise(signal, noise, snr);
```

Poměr signálu k šumu je pro každou nahrávku následující:

- bez šumu
- SNR = 4
- SNR = 7
- SNR = 10
- SNR = 15
- SNR = 20
- SNR = 30

Poté se na takto upravené nahrávky řeči a šumu aplikoval detektor řečové aktivity a postupně byly měněny parametry detektoru. Výstupem detektoru je sled jedniček a nul, symbolizujících úseky řeči a úseky pauzy. Výsledek byl bitově porovnán s referenčním výstupem VAD.

```
for i = 1 : CellCount  %CellCount je velikost nahrávky
    if referencni_vystup(i) ~= vystup_detektoru(i)
        ERR = ERR + 1; %počítadlo odlišných bitů
```

Nejlepší výsledky byly zaznamenány. Po provedení série testů bylo nutné filtrovat výsledky převyšující korekční hranici, aby nedocházelo ke zkreslení výsledků. Tato hranice je dána jako dvojnásobek směrodatné odchylky

$$kh_{ERR} = 2\sigma_{ERR}, \tag{4.1}$$

kde σ_{ERR} představuje směrodatnou odchylku celkového chybného rozhodnutí detektoru. Samotný výsledek přesnosti detekce v závislosti na nastavení detektorů nestačí. Proto se v poslední fázi na náhodně zvolené zašuměné nahrávky použije metoda spektrálního odečítání. Tato metoda je součástí VAD. Pro správnou funkci potřebuje informaci o přítomnosti či nepřítomnosti šumu v rámci (v šumovém rámci dochází k aktualizaci parametrů). Ze záznamů zbavených šumu je vypočítán odstup signálu od šumu a porovnán. Výpočet je proveden dle následující rovnice:

	$$\mathrm{SNR[dB]} = 20 \cdot \log 10\left(\frac{\mathrm{s_{zvýr}[n]}}{\mathrm{s_{zvýr}[n]} - \mathrm{s_{kont}[n]}}\right)$$	(4.2)

4.2 Databáze TIMIT

Databáze TIMIT je projektem pod záštitou americké vojenské výzkumné agentury, známé jako DARPA. Vznikala na půdách Texas Instruments, Massachusetts Institute of Technology a Stanford Research Institute v devadesátých letech. Byla vytvořena jako testovací množina řečových nahrávek pro automatické systémy vyhodnocení řeči.

TIMIT obsahuje celkem 6300 vět, 10 z nich připadá na každého z 630 řečníků z celkem 8 hlavních oblastí s rozdílnými dialekty ze Spojených států Amerických. Tab. 4.1 ukazuje rozložení řečníků podle pohlaví v jednotlivých regionech.

Každá nahrávka sestává z následujících souborů:

- soubor .WRD, obsahující jednotlivá slova a jejich vymezení v nahrávce
- soubor .PHN, obsahující časovou fonetickou transkripci
- soubor .LAB, obsahující jednotlivá slova a jejich vymezení v nahrávce
- soubor .TXT, obsahující přepis celé nahrávky
- zvukový soubor ve formátu .WAV

TIMIT není primárně určen pro testování detektorů řečové aktivity, proto bylo třeba nevhodné nahrávky odstranit. Nejčastější chyby byly následující:

1) Nepřesně označený soubor fonetického přepisu
2) Zvukový soubor obsahoval v krátkém časovém úseku vysokou míru šumu, kterou detektor vyhodnotil jako řeč
3) Soubor fonetického přepisu obsahoval nesmyslné znaky
4) Entita vykazovala míru chybovosti větší než dvojnásobek střední hodnoty chybovosti celého souboru a proto nebyla do dalšího zpracování zahrnuta

Tab. 4.1: Tabulka rozdělení řečníků podle pohlaví a regionu

Region nářečí	Muž	Žena	Celkem
New England	31	18	49
Northern	71	31	102
North Midland	79	23	102
South Midland	69	31	100
Southern	62	36	98
New York City	30	16	46
Western	74	26	100
Army Brat	22	11	33

Při provádění testů byly použity soubory fonetického přepisu PHN. Ukázka souboru PHN je na Obr. 4.2. První číselná hodnota značí levou hranici fonému, druhá číselná hodnota pravou hranici fonému a skupina znaků představuje foném. Pauzy v signálu jsou značeny jako h#.

```
0 2560 h#
2560 4130 sh
4130 4678 iy
4678 5389 hv
5389 7000 ae
7000 8313 zh
8313 8791 ax-h
8791 9170 dcl
9170 9660 d
9660 10973 aa
10973 11720 r
11720 12519 kcl
12519 14470 s
14470 16280 ux
16280 16680 dx
```

Obr. 4.2: Fonetický přepis věty "She had a dark suit" v abecedě Arpabet

4.3 Databáze VUT

Tento soubor nahrávek vznikl v tzv. bezodrazové komoře na fakultě elektrotechniky a telekomunikačních technologií, Vysokého učení technického. Všechny nahrávky byly ručně označené v programu Praat. Struktura databáze je zobrazena na oObr. 4.3 Každá podsložka řečníka obsahuje nahrávku ve formátu WAV a její slovní přepis v souboru s příponou TEXTGRID. Na tvorbě databáze se podílelo 16 řečníků, z toho 10 mužů a 6 žen, ve věku 20 až 25 let.

Izolovaná slova

Tato část obsahuje jednu nahrávku obsahující 100 krátkých slov. Slova byla vybrána tak, aby bylo těžké identifikovat jejich začátek a konec.

Izolované věty

Obsahuje 12 náhodně zvolených vět, např. z novinových článků.

Izolované věty s hlukovým pozadím

Jsou použity stejné věty jako v případě izolovaných vět, ale posluchači měli nasazená sluchátka s puštěným záznamem prostředí zamořeným šumem. Tento jev nasimulovat tzv. Lombardův efekt[3]. Ve výsledné nahrávce ale není tento šum obsažen.

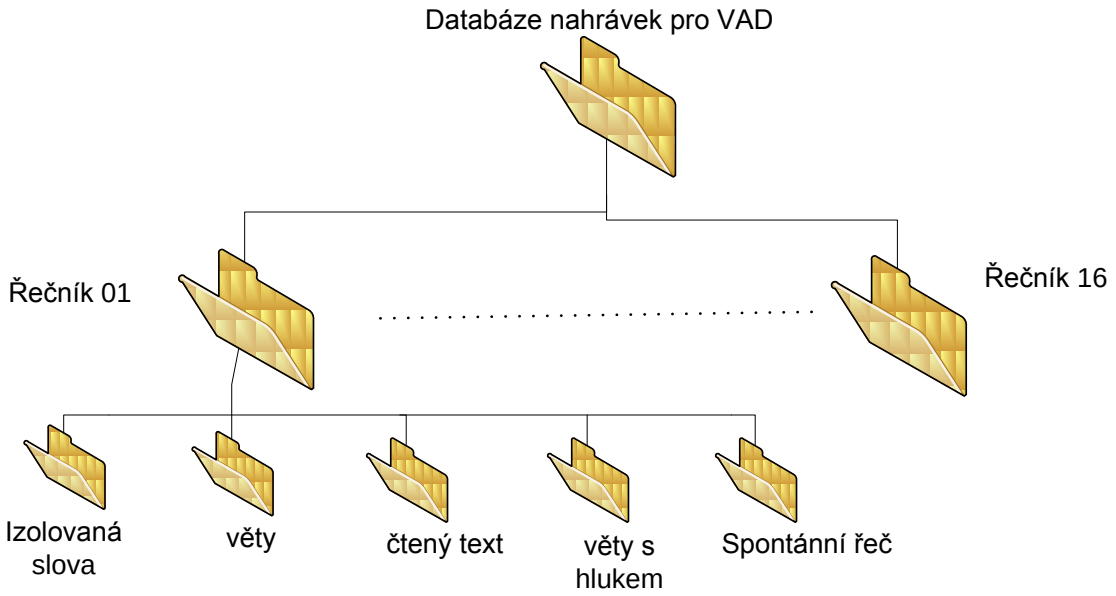
Čtený text

Celek čtený text obsahuje pouze jediný záznam četby krátké povídky. Délka nahrávky je v průměru 2 minuty a 20 sekund.

Spontánní řeč

Posluchačům byla puštěna jedenkrát puštěna krátká pohádka. Posluchači měli za úkol si pohádku zapamatovat a převyprávět vlastními slovy. Výsledné nahrávky se pohybují mezi dvěma až třemi minutami.

[3] Lombardův efekt je přirozená vlastnost mluvčích zvýšit intenzitu hlas v hlučném prostředí. Tato změna nezvýší pouze samotnou hlasitost promluvy, ale také akustické vlastnosti řeči jako jsou výška tónu nebo délka slabiky.

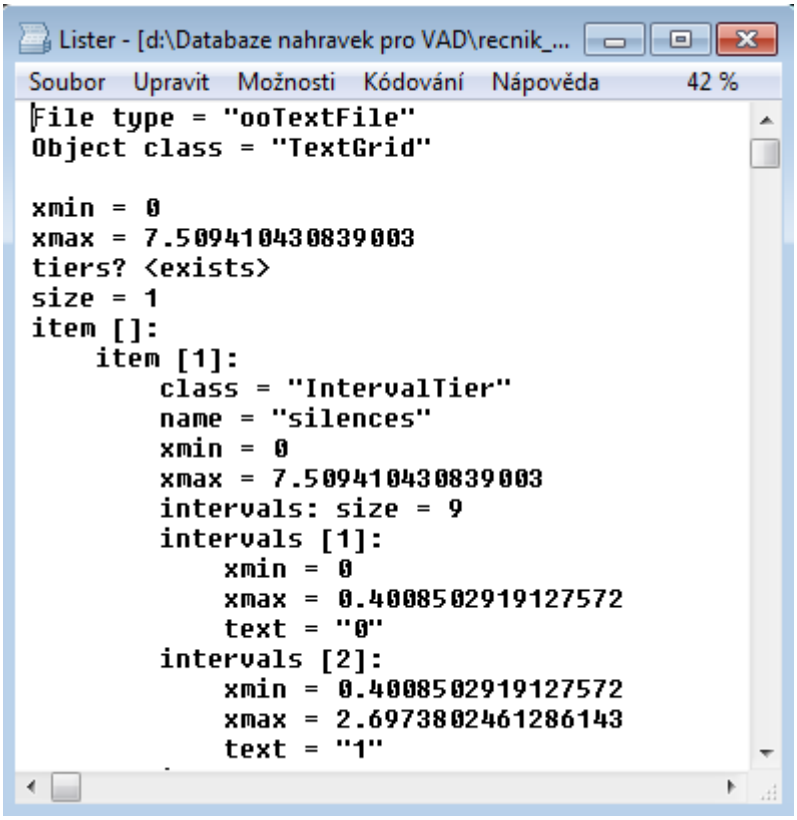


Obr. 4.3: Struktura databáze VUT

Ze souboru byly odstraněny nevhodné nahrávky, nejčastěji z důvodů:

- příliš krátký úsek šumu na začátku nahrávky pro získání prahové hodnoty
- zvukový soubor byl příliš velký a neúnosně zpomaloval program
- entita vykazovala míru chybovosti větší než dvojnásobek střední hodnoty chybovosti celého souboru a proto nebyla do dalšího zpracování zahrnuta

Díky těmto parametrům byly využity pouze nahrávky z celku „Izolované věty“ a „Izolované věty s hlukovým pozadím“. Testovací množina čítá 298 nahrávek ve formátu wav s vzorkovací frekvencí 44,1 kHz.



Obr. 4.4: Ukázka struktury souboru s příponou TEXTGRID

4.4 Testování kepstrálním detektorem

Kepstrální detektor používá dva nastavitelné parametry. Jsou jimi parametr p a parametr α.

$$p =< 0; 1 >$$

$$\alpha =< 1; 3{,}5 >$$

Parametr p určuje velikost změny střední hodnoty náhodné veličiny šumu při detekci šumového rámce a následné aktualizace prahové hodnoty. Experimentálně byl nastaven na $p = 0{,}9$.

Parametr α ovlivňuje velikost prahové hodnoty c_{th}, viz. rovnice 2.12. Při hodnotě parametru alfa rovné jedné, je prahová hodnota nejnižší a je větší šance, že bude šumový rámec vyhodnocen jako řeč. Naopak se snižuje riziko, že by mohl být řečový rámec určen jako šumový. Tato varianta by mohla nastat při hodnotě alfa = 3,5. Bylo proto potřeba najít hodnotu α takovou, aby nedocházelo ke špatnému vyhodnocení řečových rámců a zároveň nebyl zbytečně šum detekován jako řeč. K tomuto účelu byl speciálně vytvořen program v Matlabu. Parametr α byl testován v rozmezí 1-3,5 s krokem 0,1. Původně byl použit pro zpřesnění výsledků krok 0,025. Matlab ale takto

malé hodnoty zaokrouhloval a krok musel být zvýšen právě na 0,1. Díky tomu se i výpočetní čas úlohy zhruba čtyřikrát zkrátil. Zpracování jedné nahrávky trvá průměrně 50 sekund.

Ostatní nastavení kepstrálního detektoru vyplývající z povahy testovaných vzorků:

- 50% překrytí rámců
- délka segmentu 16 milisekund
- 12 rámců pro počáteční inicializaci šumu pozadí a prahové hodnoty
- Hammingovo okno
- parametr p = 0,9

Pro nalezení optimálního parametru alfa pro daný vzorek byly vypočítány míry chybovosti detektoru řečové aktivity. Výstup z detektoru byl bitově porovnán s referenčním záznamem a vznikly tři hodnoty:

- ERR (počet všech chybných rozhodnutí v záznamu)
- ERP (počet všech chybných rozhodnutí v úseku šumu)
- ERS (počet všech chybných rozhodnutí v úseku)

4.5 Testování detektorem ITU-T G.729

Detektor standardu ITU-T G.729 umožňuje nastavovat čtyři autoregresivní koeficienty podle potřeby pro optimální detekování šumu a řeči:

- AR koeficient aktualizace klouzavého průměru celkové energie β_{Ef}
- AR koeficient aktualizace klouzavého průměru úzkopásmové energie β_{El}
- AR koeficient aktualizace klouzavého průměru středního počtu průchodu nulou β_{ZC}
- AR koeficient aktualizace klouzavého průměru kmitočtu spektrálních párů β_{LSF}

Všechny AR koeficienty nabývají hodnot <0;1>. Podobně jako v případě kepstrálního VAD, program Matlab zaokrouhluje výpočty při přesnosti AR koeficientů menší jak jedna desetina. Pokud by testovací algoritmus vzal postupně kombinaci všech možností AR koeficientů s krokem 0,1, výsledkem by bylo 10^4 kombinací. Výpočetní čas by byl v tomto případě neúnosný. AR koeficienty β byly testovány od 0,1 do 0,9 s krokem 0,2. Doba testování jednoho vzorku pro všechny úrovně SNR se pohybuje v desítkách minut.

Ostatní nastavení detektoru G.729 vyplývající z povahy testovaných vzorků:

- 50% překrytí rámců
- délka segmentu 10 milisekund
- 12 rámců pro počáteční inicializaci šumu pozadí a prahové hodnoty
- váhování pomocí Hammingovo okno

Pro nalezení optimálního parametru alfa pro daný vzorek byly vypočítány míry chybovosti detektoru řečové aktivity. Výstup z detektoru byl bitově porovnán s referenčním záznamem a vznikly tři hodnoty:

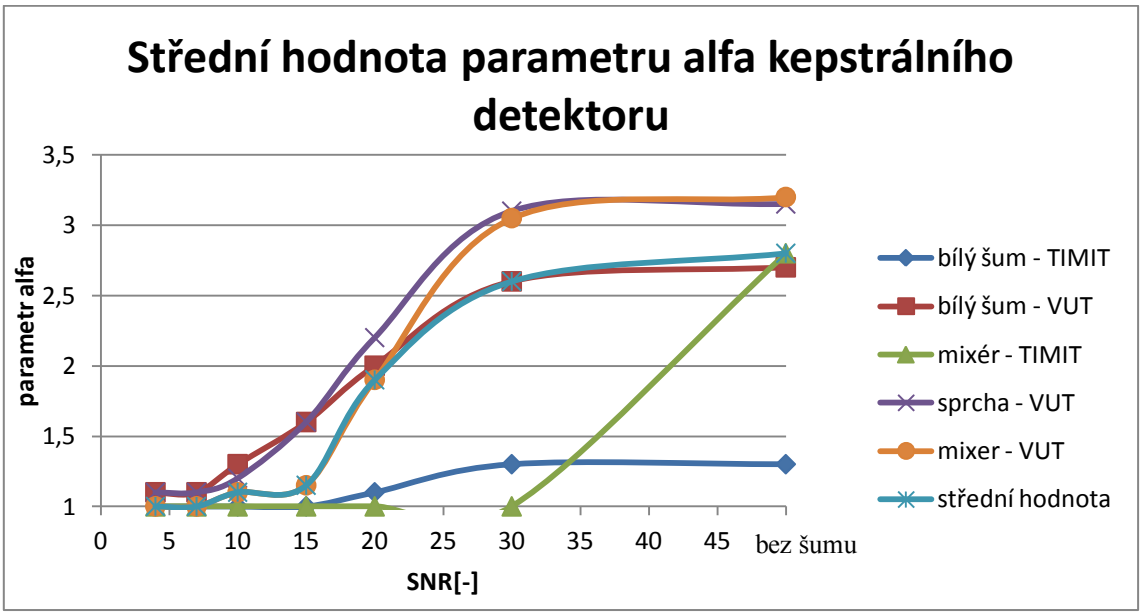
- ERR (počet všech chybných rozhodnutí v záznamu)
- ERP (počet všech chybných rozhodnutí v úseku šumu)
- ERS (počet všech chybných rozhodnutí v úseku)

4.6 Diskuze výsledků

4.6.1 Výsledky testování kepstrálním detektorem

Všechny testy byly prováděny v programovém prostřední Matlab a následně byly porovnány. Informace o počtu vzorků vyhovujících korekční hranici pro různé typy rušení a jejich úrovně jsou v Tab. 4.2. Na Obr. 4.5 je vidět chování kepstrálního detektoru pro jednotlivé typy rušení a různé úrovně šumu. Při malém odstupu signálu od šumu se hodnota parametru shoduje pro všechny typy rušení a obě testovací databáze, se zvyšujícím se poměrem SNR se zvyšuje hodnota parametr α. V grafu jsou uvedeny vždy střední hodnoty parametru α z celé testovací množiny pro dané rušení a jeho poměru k původnímu signálu. Výsledky se začínají rozcházet ne u typů rušení, ale u použitých testovacích množin. Databáze TIMIT má křivky parametru α mnohem níže položené i při vyšších SNR, než databáze VUT. Je to způsobeno značkováním referenčních vzorků. Proto byla vypočtena střední hodnota pro každou úroveň rušení a vzata jako výsledek úkolu hledání optimálního parametru α.

Z výsledků také plyne, že parametr α kepstrálního detektoru je ovlivněn hlavně tím, jak moc je signál degradovaný šumem a ne tím, o jaký typ rušení se jedná.



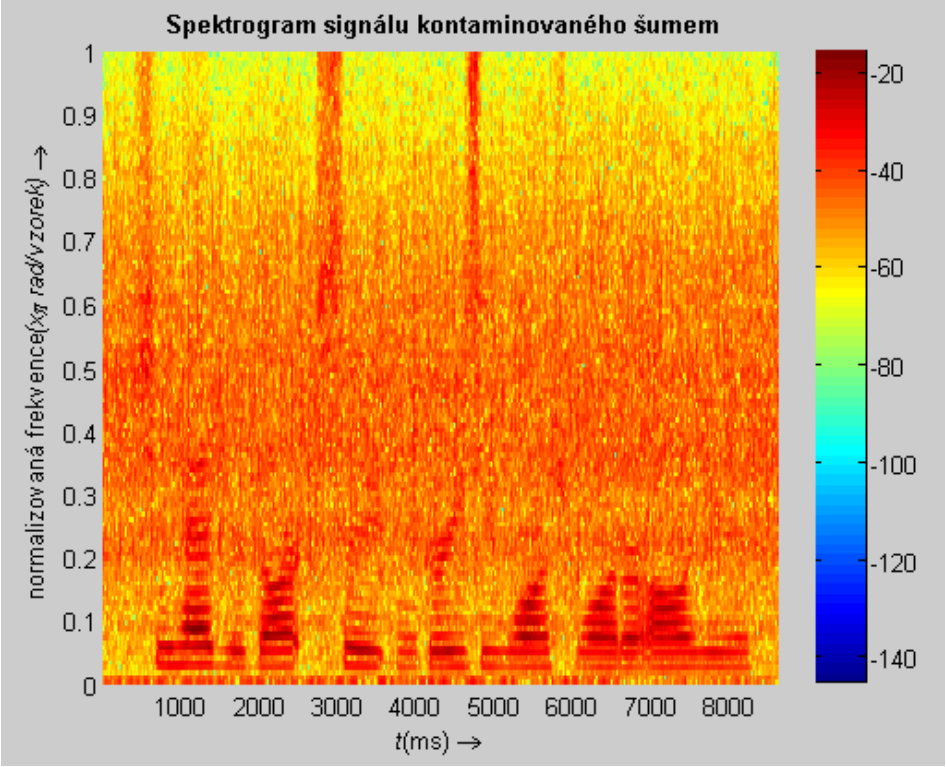
Obr. 4.5: Střední hodnoty parametru alfa pro jednotlivé typy šumu při využití databází TIMIT a VUT

O tom, jak velký vliv má parametr α na přesnost detekce šum/pauza a efektivnost filtrace při spektrálním odečítání, hovoří následující grafy. Je patrné, že čím nižší odstup signálu od šumu, tím větší budou rozdíly a naopak. Při SNR = 4 dosahuje rozdíl přesnosti detekce mezi α = 1 a α = 3,5 až **30 %**. Při aplikování metody spektrálního

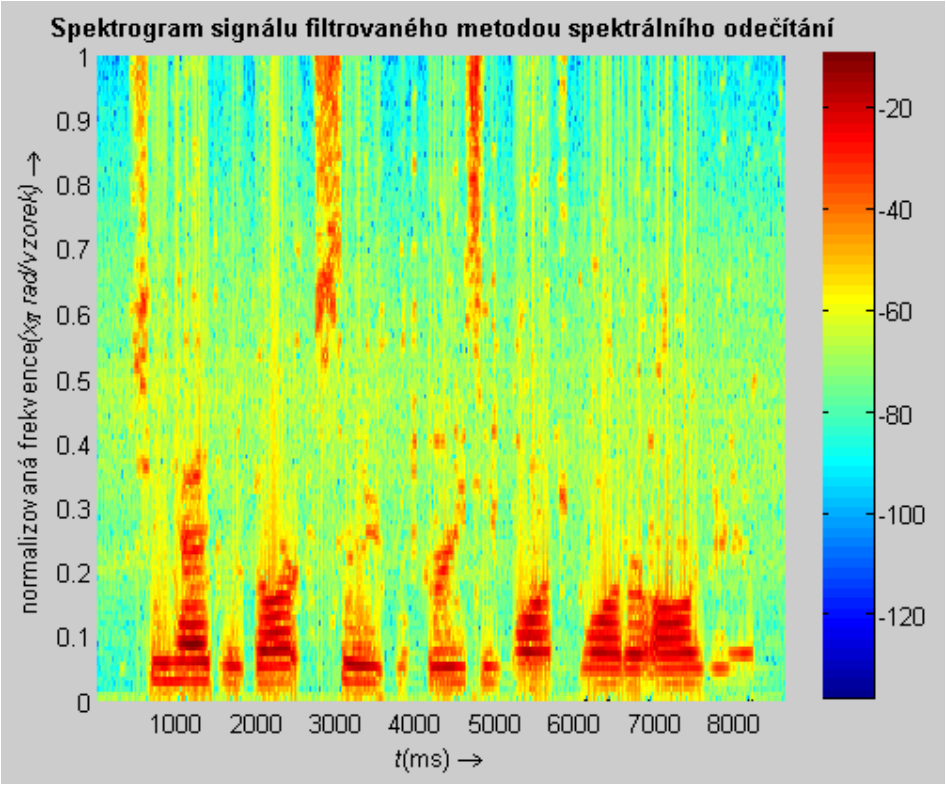
odečítání je pak rozdíl poměru ku šumu filtrovaného signálu **ΔSNR v rozmezí 0,10 až 0,50 dB, tedy zlepšení 3 až 10 %.** Naopak při degradování čistého signálu šumem o SNR = 20 se výsledky téměř shodují.

Tab. 4.2: Informace o počtu nahrávek vyhovujících korekční hranici

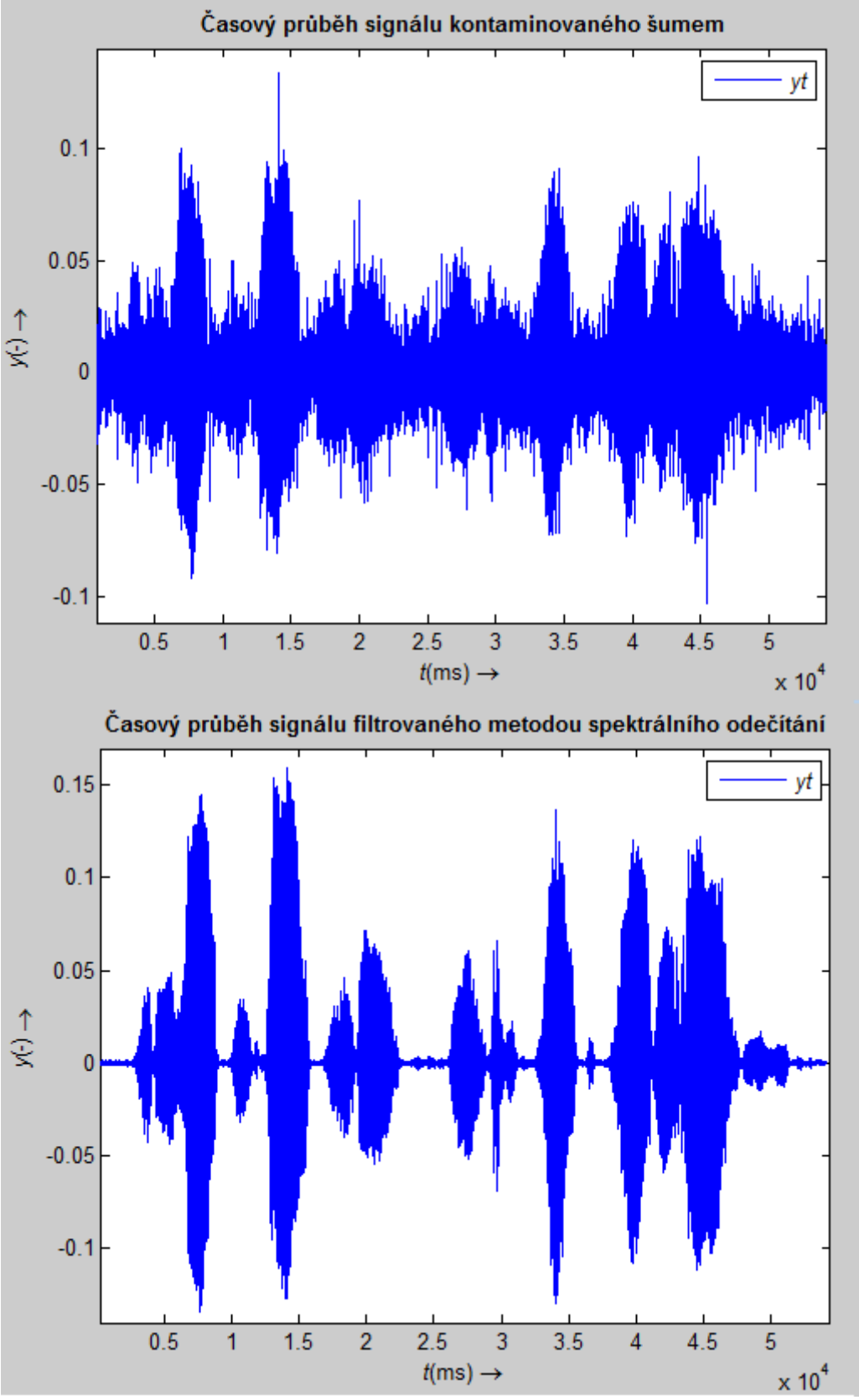
Databáze	Typ šumu	Počet nahrávek
TIMIT	bílý šum	5250
VUT	bílý šum	298
TIMIT	mixér	1909
VUT	mixér	60
VUT	sprcha	67



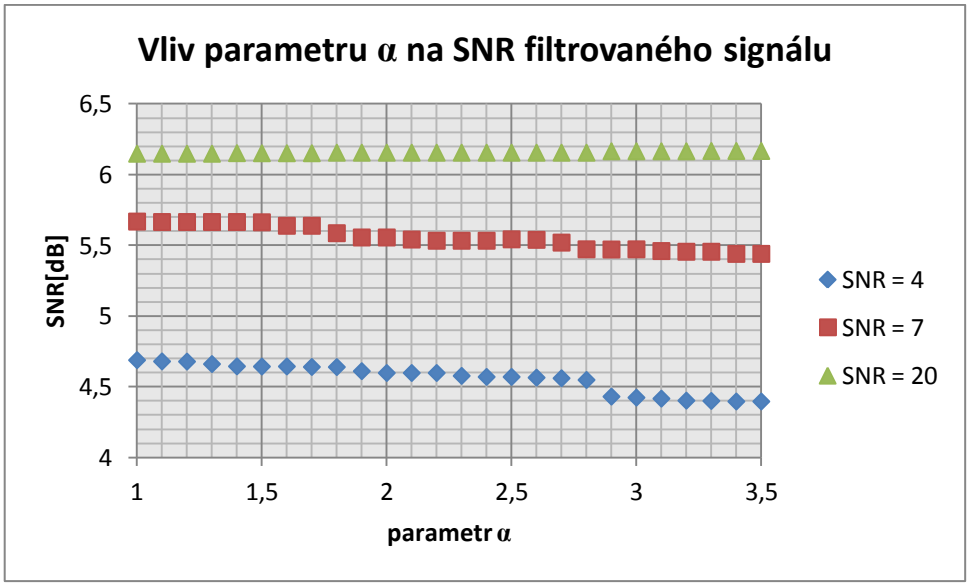
Obr. 4.6: Spektrogram původního signálu s aditivně přičteným šumem zapnuté sprchy, SNR = 4



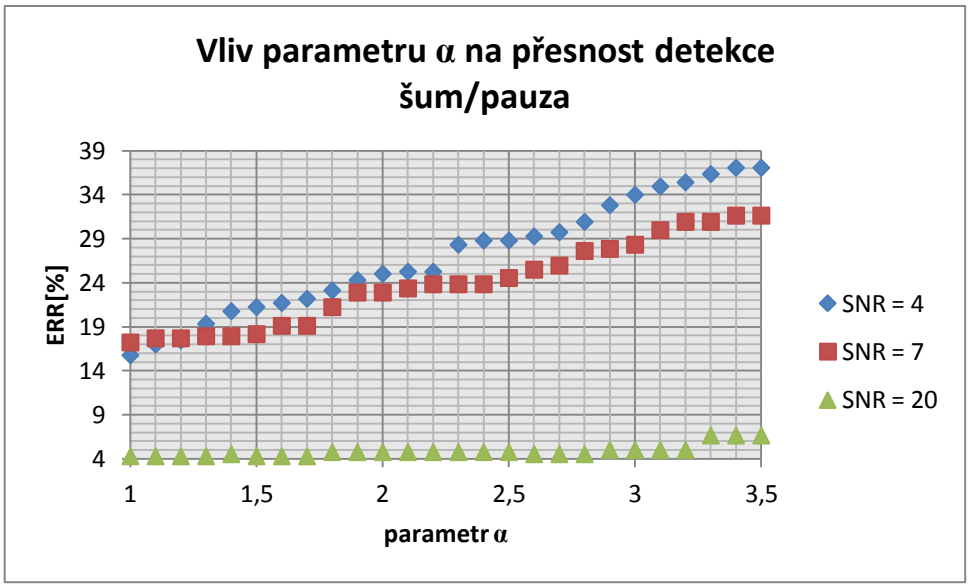
Obr. 4.7: Spektrogram signálu po odečtení šumu spektrálním odečítáním, SNR = 4,69 dB



Obr. 4.8: Nahoře: Časový průběh signálu kontaminovaného šumem zapnuté sprchy; dole: časový průběh signálu po odečtení šumu metodou spektrálního odečítání



Obr. 4.9: Vliv parametru alfa na SNR signálu filtrovaného metodou spektrálního odečítání



Obr. 4.10: Vliv parametru alfa na přesnost detekce šum/pauza signálu filtrovaného metodou spektrálního odečítání

4.6.2 Výsledky testování detektoru ITU-T G.729

Detektor G.729 byl testován pouze pro bílý šum pro obě databáze a pro šum mixéru pro databázi VUT. Informace o počtu nahrávek nepřekračujících korekční hranici je v tabulce č. Hodnoty AR koeficientů jsou závislé na povaze testovaného signálu. Pro zvukovou nahrávku namluvenou jedním řečníkem, s aditivně přičteným jedním typem rušení a stejným SNR lze v jednotlivých případech získat diametrálně odlišné optimální parametry AR koeficientů. Z toho důvodu je optimální nastavení detektoru určeno jako

průměrná hodnota výsledků jednotlivých databází. Tyto parametry dosahují hodnot 0,2 až 0,3, vyjma parametru β_{ZC}, který se pro všechny úrovně šumu blíží hodnotě 0,1. Při použití optimálních hodnot se bude spíše mluvit o kompromisních hodnotách.

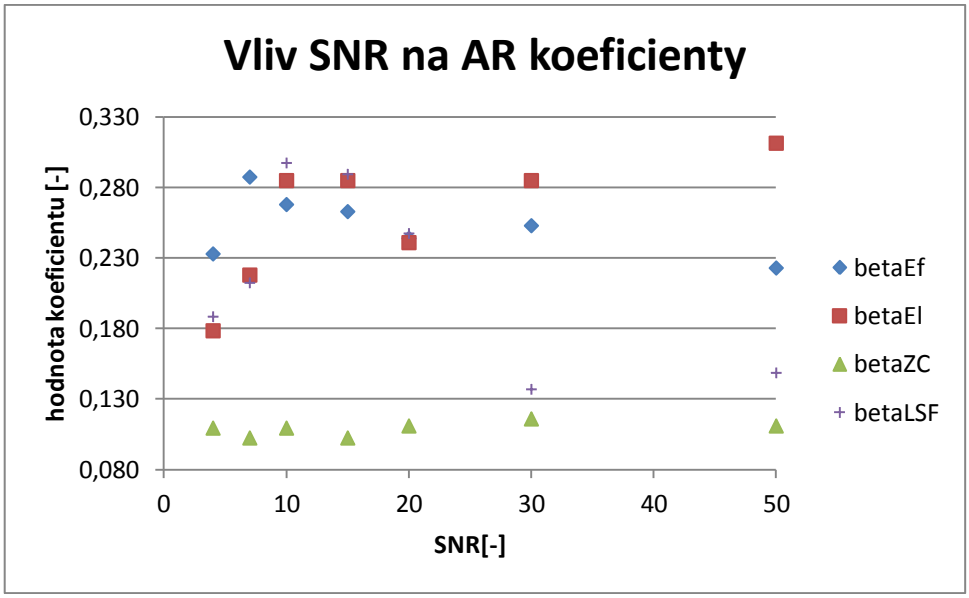
Rozdíl chyby detekce pro minimální a maximální ERR se pohybuje v jednotkách procent. Při aplikování metody spektrálního odečítání na testovacím vzorku obsahujícím aditivní šum nebyl naměřen **žádný** mezi **libovolnou kombinací** parametrů beta.

Tab. 4.3: Průměrné hodnoty AR koeficientů beta; V = databáze VUT; T = databáze TIMIT; W = bílý šum; M = šum mixéru

	SNR	4	7	10	15	20	30	bez šumu
	V-W	0,223	0,25	0,182	0,141	0,118	0,214	0,179
betaEf	T-W	0,243	0,325	0,354	0,385	0,37	0,292	0,283
	V-M	x	x	x	x	x	x	0,207
	V-W	0,205	0,264	0,291	0,355	0,441	0,241	0,244
betaEl	T-W	0,152	0,172	0,191	0,215	0,182	0,105	0,1
	V-M	x	x	x	x	x	x	0,24
	V-W	0,114	0,1	0,118	0,1	0,118	0,132	0,1
betaZC	T-W	0,105	0,105	0,101	0,105	0,104	0,1	0,1
	V-M	x	x	x	x	x	x	0,133
	V-W	0,205	0,209	0,336	0,327	0,282	0,155	0,137
betaLSF	T-W	0,172	0,216	0,259	0,252	0,213	0,119	0,11
	V-M	x	x	x	x	x	x	0,199

Tab. 4.4: Nalezené optimální hodnoty AR koeficientů beta

betaEF	**0,233**	**0,288**	**0,268**	**0,263**	**0,244**	**0,253**	**0,223**
betaEl	**0,179**	**0,218**	**0,241**	**0,285**	**0,312**	**0,173**	**0,195**
betaZC	**0,110**	**0,103**	**0,110**	**0,103**	**0,111**	**0,116**	**0,111**
betaLSF	**0,189**	**0,213**	**0,298**	**0,290**	**0,248**	**0,137**	**0,149**



Obr. 4.11: Graf vlivu poměru čistého signálu k aditivnímu šumu na AR koeficienty

Tab. 4.5: Počet nahrávek pro jednotlivé databáze a šumy, které nepřekročily korekční hranici

Databáze	Typ rušení	Počet nahrávek
TIMIT	Bílý šum	310
VUT	Bílý šum	44
VUT	Šum mixéru	60

5 ZÁVĚR

Cílem diplomové práce bylo najít efektivní metodu, který by věrohodně identifikovala a odstranila nežádoucí šum. Pro identifikaci byly zvoleny dva detektory řečové aktivity: kepstrální dvoukrokový detektor a detektor ITU-T G.729. Oba detektory dosahují velice dobrých výsledků i v silně zarušeném signálu. Oba detektory jsou závislé na správném nastavení. Pro zvýšení přesnosti detekce byly určeny optimální parametry detektorů. K tomu byly využity dvě testovací databáze, TIMIT a VUT. Databáze obsahují čistý řečový signál. K nim byly postupně aditivně přičteny pro kepstrální detektor: bílý šum, zvuk mixéru a zvuk zapnuté sprchy. Pro G.729 pouze bílý šum a zvuk mixéru. Původní plán byl otestovat všechny druhy rušení pro všechny nahrávky v databázi. Vyskytlo se ale několik překážek a komplikací, kvůli kterým bylo od tohoto záměru upuštěno a použito omezené množství nahrávek z každé databáze. Hlavním důvodem byla extrémní výpočetní náročnost, obzvlášť pro detektor G.729. Pro přesné otestování všech parametrů by vzniklo 10 000 kombinací. Při využití 7 různých úrovní šumu a několika tisíc nahrávek, což je $3{,}5*10^8$ průchodů detektorem. Jeden průchod zabere přibližně jednu sekundu, v závislosti na velikosti nahrávky. Jako kompromis byl každý parametr testován pro pět hodnot, od 0,1 po 0,9 s krokem 0,2. Tím se výpočetní doba pro jednu nahrávku snížila na 30 minut. Další problémy se týkaly softwaru Matlab, kdy docházelo k zamrzání programu, popř. náhodnému přerušení běhu testovací aplikace a nutnosti restartovat Matlab.

Pro kepstrální dvoukrokový detektor se podařilo najít optimální hodnotu parametru α. Tato hodnota se při malých odstupech signálu od šumu pro různé druhy rušení nemění. Změna nastává až u vyšších SNR, kde jsou ale rozdíly mezi výsledky s různou hodnotou parametru α minimální. Správnost měřených výsledků byla potvrzena aplikováním metody spektrálního odečítání. Rozdíl mezi nejlepším a nejhorším výsledkem dosahoval v některých případech až 10 %. Přesnost detekce se lišila až o 30 %.

Pro detektor G.729 nelze jednoznačně určit optimální nastavení. Vyplývá to z komplexity detektoru. Jednotlivé parametry jsou na sobě závislé, stejně tak jsou závislé i na testovaném vzorku. Podařilo se najít takové nastavení, aby přesnost detekce byla vždy někde uprostřed. Rozdíly v přesnosti detekce při různých nastavení detektoru se pohybují v řádu několika procent. Metoda spektrálního odečítání dosahovala pro všechna nastavení totožných výsledků. Nutno podotknout, že detektor ITU-T G.729 dosahuje lepších výsledků.

LITERATURA

[1] ATTASI, Hicham. *Metody detekce základního tónu řeči* [online]. 2008, ISSN 1213-1539 (Elektrorevue, 2008/4). Dostupné z: (http://elektrorevue.cz/cz/clanky/zpracovani-signalu/40/metody-detekce-zakladniho-tonu-reci)

[2] ITU-T. *Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70* November 1996. Dostupné z: (http://www.itu.int/rec/T-REC-G.729-199610-I!AnnB/en)

[3] ITU-T. *Coding of speech at 8 kbit/s using conjugate-structure algebric-codexited linear-prediction (CS-ACELP)* March 1996. Dostupné z: (http://itu.int/home/index.home)

[4] JOHNSON, Don. White Gaussian Noise. In: *Connexions* [online]. 2010 [cit. 2012-12-03]. Dostupné z: (http://cnx.org/content/m11281/latest/)

[5] KRČMOVÁ, M. *Fonetika* [online]. Elektronické texty. MU Brno 2003. Dostupné z: (http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html)

[6] MATLAB version 7.14.0. MATLAB R2012a. 2012. The MathWorks Inc., Masaschusetts

[7] Odfiltrování rušivých signálů ze zašumělé řeči. In: PORUBA, Jiří a Lukáš MATĚJÍČEK. *Elektrorevue* [online]. 2002 [cit. 2012-12-03]. Dostupné z: (http://www.elektrorevue.cz/clanky/02047/index.html#Kap2.1)

[8] PELIKÁN, Pavel *Databáze nahrávek pro detekci hlasové aktivity*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2011. 38 s. Vedoucí práce byl Ing. Ivan Míča

[9] POLLÁK, Karel a Petr SOVA. The study of speech/pause detectors for speech enhancemnt methods. In: *ČVUT* [online]. 1995 [cit. 2012-12-03]. Dostupné z: (http://noel.feld.cvut.cz/speechlab/publications/003_eurospeech95.pdf)

[10] PSUTKA, J. et al. *Mluvíme s počítačem česky*. Vyd. 1. Praha: Academia, 2006, 746 s. ISBN 80-200-1309-1.

[11] SIGMUND, Milan. *Analýza řečových signálů*. Vyd. 1. Brno: Fakulta elektrotechniky, 2000, 86 s. ISBN 80-214-1783-8.

[12] SMÉKAL, Z.: *Číslicové zpracování řeči (MZPR)*. Elektronické učební texty pro magisterské studium, VUT Brno, 2011.

[13] VÁCLAVÍK, Milan. *VADCRIT*. České vysoké učení technické, Fakulta elektrotechnická, 2002. Dostupné z: (http://noel.feld.cvut.cz/speechlab/start.php?page=download&lang=en)

[14] VONDRA, Martin. *Kepstrální analýza řečového signálu* [online]. 2001 (Elektrorevue, 2001). Dostupné z: (http://www.elektrorevue.cz/clanky/01048/index.html)

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

AR	Auto-Regressive, autoregresivní
ČTÚ	Český telekomunikační úřad
ČVUT	České učení technické v Praze
DARPA	Defense Advanced Research Projects Agency, agentura pro výzkum pokročilých vojenských projektů
DCT	Discrete Cosine Transform, diskrétní kosinová transformace
DFT	Discrete Fourier Transformation, diskrétní Fourierova transformace
E_f	Celková energie
EGG	Electroglottograph, elektroglotograf
E_l	Úzkopásmová energie
E_{min}	Minimální energie
ERP	Error in Pause, chybná rozhodnutí v pauze
ERR	Error Decision, celková chybná rozhodnutí
ERS	Error in Speech, chybná rozhodnutí v řeči
f	Signál v časové oblasti
F_0	Základní tón řeči
F_1,F_2	Základní formanty
FFT	Fast Fourier Transformation, rychlá Fourierova transformace
FIR	Finite Impulse Response, konečná impulzní odezva
f_v	Vzorkovací frekvence
ICA	Independent Component Analysis, analýza nezávislých komponent
LP	Linear Prediction, lineární predikce
LPC	Linear Predictive Coding, lineární prediktivní kódování
LSF	Linear Spectral Frequency, kmitočet spektrálních párů
MFCC	Mel-frequency cepstrum coefficients, melovské kepstrální koeficienty
PCM	Pulse Code Modulation, pulzní kódová modulace
RASTA	Relative Spectral
SCA	Simultaneous Component Analysis

SNR	Signal to Noise Ratio, odstup signálu od šumu
VAD	Voice Activity Detector, detektor řečové aktivity
VOIP	Voice over IP, technologie přenosu hlasu v IP sítích
ZC	Zero-crossing rate, střední počet průchodů nulovou úrovní

SEZNAM PŘÍLOH

A PŘÍLOHA PRVNÍ

A.1 Tabulka konstant detektoru G.729

Tab. A.1: Seznam konstant a jejich hodnot detektoru ITU-T G.729[4]

Konstanta	Hodnota	Konstanta	Hodnota
N_i	32	N_1	4
N_0	128	N_2	10
K_0	0	T_1	671088640**
K_1	-53687091**	T_2	738197504**
K_2	-67108864**	T_3	26843546**
K_3	-93952410**	T_4	40265318**
K_4	-134217728**	T_5	40265318**
K_5	-161061274**	T_6	40265318**
a_1	23488*	b_1	28521*
a_2	-30504*	b_2	19446*
a_3	-32768*	b_3	-32768*
a_4	26214*	b_4	-19661*
a_5	0	b_5	-30802*
a_6	28160*	b_6	-19661*
a_7	0	b_7	30199*
a_8	16384*	b_8	-22938*
a_9	-19065*	b_9	-31576*
a_{10}	0	b_{10}	-17367*
a_{11}	22400*	b_{11}	-27034*
a_{12}	30427*	b_{12}	29959*
a_{13}	-24576*	b_{13}	-29491*
a_{14}	23406*	b_{14}	-28087*

[4] Hodnoty označené * musí být vyděleny číslem 2^{15}; hodnoty označené ** musí být vyděleny číslem 2^{31}

A.2 Obsah přiloženého CD

Přiložené CD obsahuje 3 adresáře:

Adresář **„Funkce“** obsahuje všechny m-funkce využité při zpracování diplomové práce.

Adresář **„Zvuky“** obsahuje náhodně zvolené nahrávky pro ověření funkce detektoru řečové aktivity a jejich označené ekvivalenty ve formátu PHN a Textgrid.

Adresář **„Šumy“** obsahuje šumové nahrávky ve formátu WAV.

V kořenovém adresáři CD je uložena diplomová práce ve formátu PDF a všechny výsledky obdržené testováním v souboru XLS.

Aplikace se spustí v programu Matlab pomocí příkazu `guide GUI`. Načte se editační okno grafického rozhraní. Stiskem kláves CTRL+T se aplikace aktivuje. Ovládání aplikace je intuitivní.